Mô tả:  
Walmart, một trong những cửa hàng bán lẻ hàng đầu tại Hoa Kỳ, muốn dự đoán doanh số bán hàng và nhu cầu một cách chính xác. Có những sự kiện và ngày lễ cụ thể ảnh hưởng đến doanh số bán hàng hàng ngày. Có sẵn dữ liệu bán hàng cho 45 cửa hàng của Walmart. Doanh nghiệp đang đối mặt với thách thức do nhu cầu không thể dự đoán và đôi khi hết hàng, do sử dụng thuật toán học máy không phù hợp. Một thuật toán ML lý tưởng sẽ dự đoán nhu cầu một cách chính xác và xem xét các yếu tố như điều kiện kinh tế bao gồm CPI, chỉ số thất nghiệp, v.v.

Walmart thực hiện nhiều sự kiện giảm giá quảng cáo trong suốt năm. Những sự giảm giá này tiên phong cho những ngày lễ nổi tiếng, bốn ngày lễ lớn nhất trong số đó là Super Bowl, Lễ lao động, Lễ tạ ơn và Giáng sinh. Các tuần bao gồm những ngày lễ này được đánh giá cao hơn năm lần so với các tuần không phải là ngày lễ. Một phần thách thức của cuộc thi này là mô hình hóa tác động của việc giảm giá trong những tuần lễ ngày lễ này trong trường hợp không có đầy đủ/dữ liệu lịch sử lý tưởng. Dữ liệu bán hàng lịch sử cho 45 cửa hàng Walmart nằm ở các khu vực khác nhau có sẵn.

Mục tiêu:

* Hiểu về Bộ dữ liệu.
* Phát hiện dấu hiệu bất thường.
* Xây dựng các mô hình hồi quy để dự đoán doanh số bán hàng liên quan đến một hoặc nhiều thuộc tính.
* Đánh giá các mô hình và so sánh các điểm số tương ứng như R2, RMSE, v.v.

1. Hiểu về bộ dữ liệu

A screenshot of a computer

Description automatically generated

Dữ liệu cột bao gồm:

17 cột

421570 dòng

Ý nghĩa của các cột:

| Cột | Ý Nghĩa |
| --- | --- |
| Store | Đây là một định danh số học cho chuỗi cửa hàng |
| Date | Ngày ghi nhận dữ liệu. |
| Isholyday | Một giá trị nhị phân chỉ ra liệu tuần ghi nhận có là tuần nghỉ lễ (1) hay không (0). |
| Dept | Đây là một định danh số học cho các phòng ban trong cửa hàng. |
| Weekly\_Sales | Doanh bán hàng tuần được ghi nhận cho một phòng ban cụ thể trong cửa hàng. |
| Temperature | Nhiệt độ ghi nhận vào ngày cụ thể. |
| Fuel\_Price | Giá nhiên liệu vào ngày cụ thể. |
| Markdown1 to Mardown5 | Các loại khuyến mãi (chiết khấu) khác nhau. |
| CPI | Chỉ số giá tiêu dùng (Consumer Price Index). |
| Unemployment | Tỷ lệ thất nghiệp. |
| Type | Đây là một định danh số học cho loại cửa hàng. |
| Size | Kích thước của cửa hàng |

1. EDA
   1. Kiểm tra dữ liệu

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

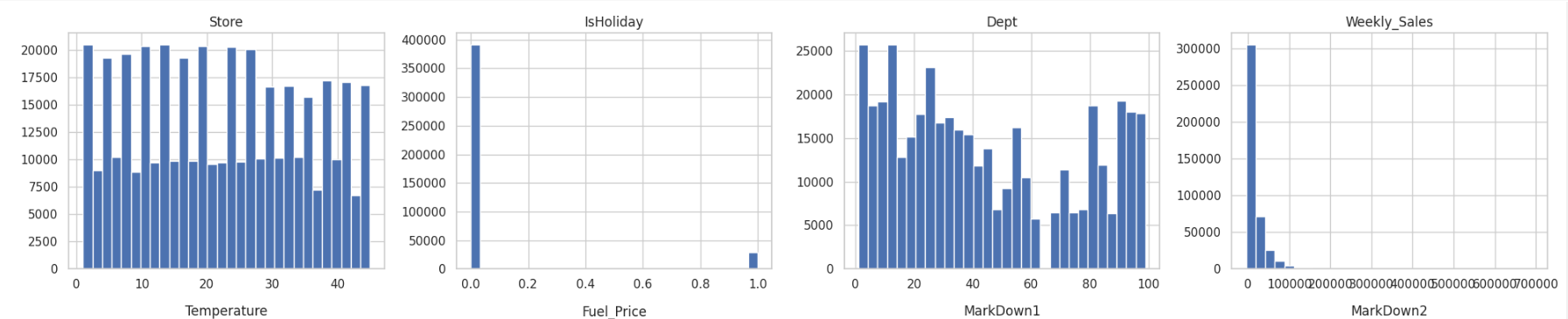
A screenshot of a computer screen

Description automatically generated

A group of colorful text boxes

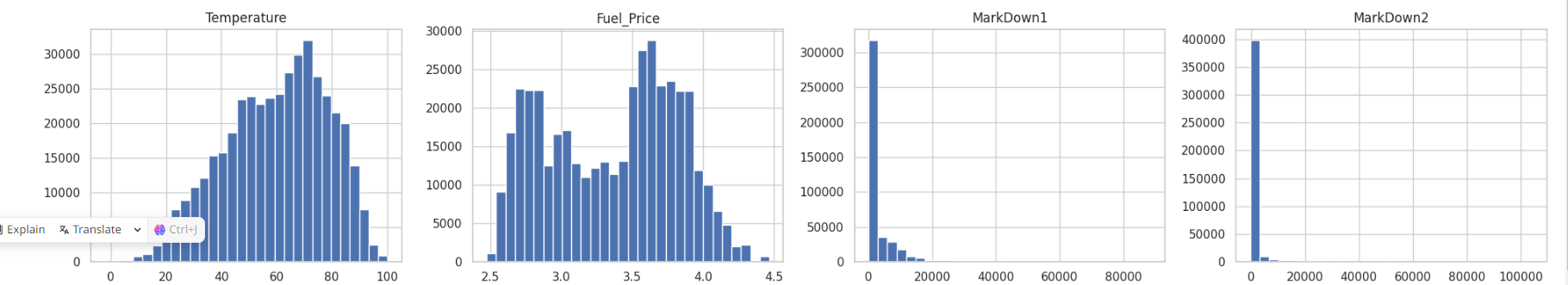
Description automatically generated with medium confidence

* 1. Phân tích và đánh giá sơ bộ



IsHoliday: Đây là một biến nhị phân chỉ ra liệu một tuần có phải là tuần lễ nghỉ hay không. Dự kiến số tuần nghỉ lễ sẽ ít hơn số tuần không phải nghỉ lễ, vì lễ nghỉ xảy ra tương đối ít.

Weekly\_Sales: Phân bố của cột Weekly\_Sales có đuôi phải dày đặc. Điều này cho thấy số lượng bán hàng cao hơn không phổ biến, trong khi số lượng bán hàng thấp hơn xảy ra thường xuyên.



Fuel\_Price: Phân bố của cột Fuel\_Price có hơi lệch phải. Điều này cho thấy giá nhiên liệu cao hơn không phổ biến và giá thấp hơn thường xuyên được quan sát.

A graph with numbers and a bar

Description automatically generated with medium confidence

CPI, Temperature và Unemployment: Các cột này có phân bố gần như chuẩn. Phân phối của Temperature có hơi lệch trái, cho thấy nhiệt độ thấp có thể phổ biến hơn. CPI (Chỉ số giá tiêu dùng) và Unemployment có phân phối tương đối đối xứng.

Các cột MarkDown1 đến MarkDown5: Những cột này có sự chênh lệch lớn về giá trị, đa phần là 0, cho thấy các chiết khấu hoặc khuyến mãi không được áp dụng hàng tuần. Điều này cho thấy chỉ có một số tuần cụ thể có sự xuất hiện của các chiết khấu, trong khi trong phần lớn các tuần khác, các giá trị này vẫn giữ ở mức 0.

A graph with blue lines

Description automatically generated

Size và Type: Quan sát cho thấy cỡ cửa hàng lớn hơn (phản ánh bởi cột "Size") và cửa hàng thuộc loại 3 (phản ánh bởi cột "Type") phổ biến hơn trong tập dữ liệu. Điều này có thể cho thấy phần lớn cửa hàng trong tập dữ liệu là cỡ lớn và thuộc loại 3

A graph showing a sales chart

Description automatically generated with medium confidence

Biểu đồ trên cho thấy Weekly\_Sales theo thời gian. Có một mẫu rõ ràng của doanh số bán hàng tăng cao vào những thời điểm cụ thể trong năm, có thể tương ứng với mùa mua sắm sôi động như kỳ nghỉ cuối năm.

1. Lựa chọn mô hình và kết quả
2. Xem sự tương quan

A screenshot of a graph

Description automatically generated

Weekly\_Sales và Size: Có một mối tương quan tích cực giữa Weekly\_Sales và Size, cho thấy các cửa hàng lớn có xu hướng có doanh số bán hàng cao hơn. Điều này cho thấy kích thước cửa hàng có thể ảnh hưởng đến hiệu suất bán hàng.

Weekly\_Sales và Dept: Weekly\_Sales cũng có mối tương quan tích cực với Dept, cho thấy một số bộ phận trong các cửa hàng có xu hướng có doanh số bán hàng cao hơn so với các bộ phận khác.

Weekly\_Sales và Type: Có một mối tương quan âm nhẹ giữa Weekly\_Sales và Type, cho thấy một số loại cửa hàng có xu hướng có doanh số bán hàng thấp hơn. Điều này cho thấy loại cửa hàng có thể ảnh hưởng đến doanh số bán hàng.

Weekly\_Sales và các đặc trưng MarkDown: Các đặc trưng MarkDown không có mối tương quan đáng kể với Weekly\_Sales. Điều này cho thấy các giảm giá quảng cáo này có thể không ảnh hưởng đáng kể đến doanh số bán hàng tổng thể.

Fuel\_Price và MarkDown1, MarkDown4 và MarkDown5: Fuel\_Price có mối tương quan tích cực với MarkDown1, MarkDown4 và MarkDown5. Điều này cho thấy các giảm giá quảng cáo cụ thể này có thể liên quan đến giá nhiên liệu một cách nào đó.

CPI và Unemployment: CPI và Unemployment có mối tương quan âm. Điều này là điều dễ hiểu, vì tỷ lệ thất nghiệp cao thường dẫn đến giá trị chỉ số giá tiêu dùng thấp hơn. Mối tương quan này cho thấy một mối quan hệ nghịch đảo giữa giá tiêu dùng và tỷ lệ thất nghiệp.

Type và Size: Type có mối tương quan âm mạnh với Size, cho thấy loại cửa hàng liên quan đến kích thước của nó. Điều này cho thấy các loại cửa hàng khác nhau có tend to có kích thước khác nhau.

Nhiệt độ và các biến khác: Nhiệt độ không có mối tương quan mạnh với các biến khác được đề cập. Điều này cho thấy nhiệt độ có thể không đóng vai trò quan trọng trong việc ảnh hưởng đến doanh số bán hàng hoặc các yếu tố khác được đề cập.

1. Chọn biến

Việc lấy biến "Size\_Type" vào có thể là một cách để tạo ra một biến kết hợp từ hai biến "Size" và "Type". Việc này có thể giúp mô hình học máy của bạn hiểu rõ hơn về mối quan hệ giữa kích thước của cửa hàng và loại cửa hàng đối với doanh số bán hàng.



A close-up of a computer screen

Description automatically generated

A screenshot of a computer program

Description automatically generated

1. Mô hình sử dụng
   * Linear Regression
   * DecisionTreeRegressor(max\_depth: 7),
   * Random Forest (n\_estimators=100, random\_state=42)
   * Extreme Gradient Boosting (n\_estimators=20,max\_depth=9 )
2. Kết quả

A table with numbers and letters

Description automatically generated

1. **Kết luận và đánh giá**

Dựa trên bảng trên, ta có thể rút ra các đánh giá và kết luận như sau:

* Linear Regression: Mô hình Linear Regression có MAE cao (16009.35) và RMSE lớn (23864.18), cho thấy độ chính xác dự đoán của mô hình không tốt. R2 Score cũng rất thấp (0.09), chỉ ra rằng mô hình không tốt trong việc giải thích phương sai của dữ liệu.
* Decision Tree: Mô hình Decision Tree có MAE (6735.31) và RMSE (11271.91) tương đối thấp, cho thấy độ chính xác dự đoán tốt hơn so với Linear Regression. R2 Score (0.8) cũng cao hơn, cho thấy mô hình Decision Tree khớp tốt hơn với dữ liệu huấn luyện.
* Random Forest: Mô hình Random Forest có MAE rất thấp (1793.8) và RMSE (4618.48) cũng thấp, chỉ ra rằng mô hình có khả năng dự đoán chính xác và có độ chính xác cao hơn so với hai mô hình trước. R2 Score (0.97) cũng rất cao, cho thấy mô hình Random Forest khớp rất tốt với dữ liệu huấn luyện.
* Extreme Gradient Boosting: Mô hình Extreme Gradient Boosting cũng có MAE khá thấp (2197.62) và RMSE (4248.02) tương đối thấp. R2 Score (0.97) cho thấy mô hình có khả năng dự đoán tốt và khớp tốt với dữ liệu huấn luyện.

Tổng quát, mô hình Random Forest và Extreme Gradient Boosting có hiệu suất tốt hơn so với Linear Regression và Decision Tree. Random Forest có hiệu suất tốt nhất với MAE và RMSE thấp nhất, cùng với R2 Score cao nhất, chỉ ra rằng mô hình này có khả năng dự đoán tốt nhất trong số các mô hình được đánh giá.