KPIM
Analyzing For Future

KPIM DATATHON

Bài toán thương mại điện tử

KPIM DATATHON:

Phân tích dữ liệu vận chuyển đơn hàng thương mại điện tử

A. Các thông tin quan trọng về dự án

1. Phạm vi kiến thức và công cụ cần sử dụng

Dự án yêu cầu thí sinh áp dụng kinh nghiệm và kĩ năng về phân tích dữ liệu với data về Logistic để trả lời các câu hỏi liên quan đến việc vận chuyển hàng hóa và dự báo, tối ưu hóa hệ thống phân phối hàng hóa giúp đảm bảo tiến độ giao hàng. Trong đó, thí sinh có thể lựa chọn đa dạng các công cụ khác nhau và không giới hạn bởi:

a. Ứng dụng ngôn ngữ SQL để truy vấn dữ liệu và ETL dữ liệu
b. Ứng dụng bảng tính Excel để phân tích, tạo biểu đồ hoặc tạo mô hình thống kê
c. Ứng dụng công cụ PowerBI để trực quan hóa báo cáo
d. Ứng dụng công cụ thống kê R để phân tích chuyên sâu
e. Áp dụng Machine Learning với ngôn ngữ lập trình Python
f. Tổng hợp thông tin phân tích và thuyết trình trên công cụ Slides (PowerPoint)

Từ dữ liệu ở các nguồn khác nhau được cung cấp, thí sinh sẽ cần sử dụng công cụ lưu trữ hoặc đồng bộ dữ liệu về chung 1 nơi. Từ đó thiết lập mô hình phân tích hợp lý và tính toán các chỉ số cần thực hiện để phát hiện những thông tin từ yêu cầu. Áp dụng bổ sung các công cụ thống kê và thuyết trình về dự án trên báo cáo hoặc các slide với biểu đồ.

2. Giới thiệu chung về dự án

KPIM Ecommerce là một công ty thương mại điện tử cung cấp cho khách hàng nhiều loại sản phẩm đa dạng. Thị trường của KPIM Ecommerce trải rộng khắp các tỉnh thành toàn quốc. Công ty vận chuyển hàng triệu đơn hàng mỗi năm để nâng cao chất lượng đời sống của khách hàng. Tiêu chí số một của KPIM Ecommerce là luôn luôn đáp ứng khách hàng vượt trên sự mong đợi của họ và một trong những tiêu chí cần đáp ứng đó là tốc độ giao hàng. Để xây dựng một thương hiệu thương mại điện tử tốt thì việc nhanh chóng quay vòng trong khâu vận chuyển hàng hóa cho khách hàng vô cùng quan trọng. Tuy nhiên, việc quay vòng càng nhanh thì mạng lưới hậu cần/giao hàng càng rộng hoặc càng tốn kém. Chính vì thế, việc phân tích dữ liệu để cân bằng chi phí giao hàng với đảm bảo chất lượng và tốc độ giao hàng là điều mà KPIM Ecommerce luôn luôn quan tâm.

B. Thử thách phân tích

Thử thách của bạn là phân tích địa lý của khách hàng, mạng lưới phân phối, tổ hợp sản phẩm và thói quen mua sắm của khách hàng của KPIM Ecommerce nhằm mục đích trực quan hoá các thông tin hữu ích để công ty thấy được sự liên quan giữa tốc độ phân phối sản phẩm/gói hàng và lòng trung thành của khách hàng

1. Các câu hỏi phân tích

a. Mạng lưới phân phối hiện tại có đáp ứng được nhu cầu của khách hàng trên các tỉnh thành không? Có trường hợp sản phẩm nào đặc biệt cụ thể không?
b. Những điểm bất cập trong mạng lưới phân phối hiện tại là gì? KPIM Ecommerce cần phải thay đổi và tối ưu hóa mạng lưới phân phối thế nào để làm việc hiệu quả?
c. Có sản phẩm hoặc danh mục sản phẩm cụ thể nào cần được chuẩn bị sẵn hàng tồn không? Nếu có thì vị trí kho và khu vực phân phối tồn nên ở đâu?
d. Khách hàng nhận được sản phẩm đúng thời hạn có mua sắm nhiều hơn so việc gửi hàng trễ hay không? Tìm hiểu mối quan hệ giữa hiệu suất giao hàng và doanh thu
e. Dự đoán xu hướng và số đơn hàng tương lai cũng như tỷ lệ đơn hàng bị trễ.
f. Có cách nào phân bổ lại địa điểm vận chuyển hàng theo nơi đến hoặc dự đoán từ đó phân bổ nơi lưu trữ hàng hóa thuận lợi cho việc vận chuyển nhanh hay không?

2. Dữ liệu

Dữ liệu này được trích xuất và ẩn danh từ KPIM. Tập dữ liệu lớn bao gồm nhiều thông tin khác nhau trong đó dữ liệu về đơn hàng được chia thành 2 nơi, file Excel chứa 1 phần dữ liệu dưới 1 triệu dòng và 1 phần lớn dữ liệu trong SQL Server với hơn 6 triệu dòng. Đồng thời các thông tin danh mục lưu tại GoogleSheet và CSV.

a. Dữ liệu đơn hàng trong Excel *(File KPIM Ecommerce Data 1)*
b. Dữ liệu đơn hàng trong SQL Server:
   - **Server**: 222.252.14.117
   - **Database**: datathon
   - **Table**: dbo.ecommerce_data
   - **User**: datathon
   - **Password**: KPIMDatathon2022
c. Dữ liệu kho hàng *(File txt KPIM Ecommerce Distribution Center)*
d. Dữ liệu trạng thái đơn hàng *(File csv KPIM Ecommerce Order Status)*
e. Dữ liệu các tỉnh thành Việt Nam: Link Google Sheet
f. Từ điển dữ liệu *(File Excel KPIM Ecommerce Data Dictionary)*

C. *Nội dung các yêu cầu thực hiện*

1. Phân tích mô tả và thiết kế report, dashboard (80%)

a. **Yêu cầu:** Dựa vào dữ liệu được cung cấp hãy đưa ra các phân tích và trả lời các câu hỏi đặt ra trong bài toán của Datathon. Sau khi đã có các phân tích cụ thể, hãy tổng hợp chúng bằng cách trực quan hóa trên một công cụ báo cáo và BI hoặc đưa vào các slide thuyết trình để trình bày về nội dung phân tích được.

Đơn đặt hàng (Orders):

- Các đơn hàng thường đến từ đâu? Tỷ trọng nhu cầu theo các tỉnh thành như thế nào? Top tỉnh thành có nhu cầu lớn nhất và nhỏ nhất?
- Sản phẩm nào được đặt nhiều nhất? Có sản phẩm nào được đặt nhiều tại một khu vực hoặc tỉnh thành cụ thể hay không?
- Mức độ tăng trưởng đơn hàng theo các tháng so với độ tăng về các đơn vận chuyển trễ như thế nào, có khác biệt gì lớn hay không?

Doanh thu (Revenues):

- Thống kê doanh thu theo các loại mặt hàng, mặt hàng nào có doanh thu lớn nhất?
- Thống kê xu hướng doanh thu theo ngày / tháng / quý / năm ?
- Tại sao một số thời điểm trong năm doanh thu bị thụt giảm? Vấn đề thụt giảm doanh thu có thể đến từ đâu và giải quyết chúng như thế nào?
- Tỷ trọng doah thu theo khu vực và các tỉnh? Top tỉnh thành doanh thu lớn nhất?

Vận chuyển trễ (Late Deliveries):

- Tỷ lệ % các đơn hàng hoàn thành vận chuyển? Tỷ lệ % các đơn hàng vận chuyển trễ? Tỷ lệ % trên có ảnh hưởng và khác biệt theo các nhóm hàng hoặc tỉnh thành?
- Tỷ lệ % vị trí xuất hàng và nhận hàng cùng tỉnh thành cho các đơn hàng trễ. Tỷ lệ % vị trí xuất hàng và nhận hàng khác khu vực vùng miền của các đơn hàng trễ.
- Có sự liên quan gì giữa khoảng cách vận chuyển đơn hàng và thời gian vận chuyển? Mối liên quan trên có đặc biệt bởi các mặt hàng nào cụ thể hay không?

Thời gian vận chuyển (Delivery Lead-time):

- Trung bình khoảng thời gian vận chuyển các đơn hàng? Phân bổ theo nhóm ngày hoặc độ trễ? Phân bổ chỉ số đó theo vị trí xuất hàng và sản phẩm như thế nào?
- Vị trí phân bổ các kho hàng và tỷ trọng trên các tỉnh thành là bao nhiêu? Top kho hàng lớn chiếm tỷ trọng lớn trên tổng số đơn hàng vận chuyển?

2. Phân tích thống kê nâng cao

a. **Yêu cầu**: Phân tích thống kê dựa vào các câu hỏi và phân tích mô tả ban đầu để giải quyết 2 bài toán lớn nhất với KPIM Ecommerce: **dự báo đơn hàng mới** và **tối ưu hóa vị trí kho hàng và đường vận chuyển hàng hóa**. Từ việc áp dụng các mô hình thống kê, giải pháp máy học hoặc ứng dụng thuật toán đưa ra các gợi ý để giảm thời lượng vận chuyển hàng, gia tăng niềm tin của khách hàng và đơn hàng / doanh thu.

Có thể cân nhắc áp dụng một số mô hình và thuật toán, công cụ sau:

Xử lý dữ liệu: Giản lược cấp độ chi tiết của dữ liệu (Data Granularity) theo các chiều phân tích và các chỉ số phân tích trong mô hình thống kê (Product Category, Location, Customer, Date/Time, …) để dễ dàng áp dụng mô hình phân tích và gia tăng tốc độ tính toán, thống kê ***(****Gợi ý: sử dụng SQL)*

Tóm tắt thống kê: Thực hiện tính toán thống kê cơ bản và tóm tắt các yếu tố trong mô hình kèm theo là các parameter cơ bản thể hiện dữ liệu cũng như thống kê cơ bản của dữ liệu để hiểu về từng đối tượng ảnh hưởng.

Sử dụng mô hình dự báo: Áp dụng các mô hình dự báo hợp lý theo thời gian như (ARIMA, Exponential Smoothing, Random Forest, XGBoost, …)