* **Các kiến thức chuyên môn**
* Các loại chart, khi nào sử dụng các chart nào.
* Quy trình xử lý và làm sạch dữ liệu:

+ Thu thập dữ liệu

+ Khai phá dữ liệu (EDA)

+ Xác định và xử lý Missing value

+ Xử lý dữ liệu trùng lặp

+ Xử lý outlier

+ Chuẩn hoá dữ liệu

* Các cách phát hiện outlier và xử lý

+ Với dữ liệu numerical

+ Với dữ liệu categorical

+ Một số thuật toán phát hiện outlier: Kmean, HDBSCAN, Isolation forest, Local Outlier Factor, SVM one class

* Phân tích hồi quy
* Phân tích tương quan
* Phân tích phân cụm
* Các phương pháp kiểm định mô hình
* Kiểm định giả thuyết thống kê
* Quy trình phân tích dữ liệu

+ Bussiness understanding

+ Data understanding: Theme, data mechanism, other (sourcce, null, đáng tin, đúng định dạng, make sense

+ data preparation: Gather -> Discover - > Clean -> transform -> enrich -> store

+ Modeling / Analysis

+ Evaluation (Nếu chưa ổn thì quay lại Business understanding)

+ Deployment/ Execution

* Mô hình phân tích đa chiều- OLAP

+ Quy trình xây dựng OLAP: Nghiệp vụ -> Xác định dim, fact -> Voi dim -> Data model logic, data model vật lý -> Phát triển các truy vấn và báo cáo

+ Phân tích với OLAP: Drill- down, Rollup, data slicing, data dicing

* Quy trình xây dựng một hệ thống
* Các sơ đồ UML: Sơ đồ phân cấp chức năng, ERD, sơ đồ usecase, sơ đồ lớp, sơ đồ tuần tự, sơ đồ hoạt động
* ETL
* Tối ưu query

**Trả lời:**

1. Các loại chart. Khi nào thì sử dụng chart nào?

* Bar chart: So sánh/ Biểu diễn dữ liệu phân loại thành các nhóm riêng biệt (thường là biểu diễn dữ liệu categorical)
* Biểu đồ cột chồng: Biểu đồ này cho thấy sự đóng góp của các ngăn riêng lẻ tạo nên các cột
* Biểu đồ cột kết hợp đường: Khi có hỗn hợp các loại chuỗi dữ liệu. Hoặc thể hiện chuỗi dữ liệu bổ sung cùng với dữ liệu chính
* Biểu đồ tròn: Để biểu diễn mức đóng góp của các phần trong tổng thể
* Donut chart: Giống với biểu đồ tròn. Biểu diễn đóng góp của các phần vào tổng thể
* Radar chart: Để nghiên cứu so sánh các chuỗi dữ liệu khác nhau
* Funnel chart: Thể hiện dòng/ giảm luỹ tiến của một số liệu qua các giai đoạn
* Biểu đồ đường: Trực quan hoá xu hướng của các chuỗi dữ liệu trong các khoảng thời gian
* Scatter plot: Biểu diễn phân tán của dữ liệu. Biểu đồ này được sử dụng để so sánh giữa hai trục số không giống nhau, trong đó một trục không bao giờ là số
* Biểu đồ vùng: Là phần che phủ dưới các đường, giúp dễ dàng so sánh các mức dữ liệu. Biểu đồ này chủ yếu hữu ích để nhấn mạnh sự thay đổi về số liệu theo thời gian.
* Biểu đồ xếp chồng (stacked area): Biểu đồ vùng xếp chồng cho thấy mối quan hệ của các phần với tổng thể. Xem các ngăn xếp hoặc các yếu tố riêng lẻ đã đóng góp bao nhiêu vào tổng giá trị theo thời gian
* Biểu đồ vòng (dial chart): Chỉ ra giá trị hiện tại trong một phạm vi
* Bubble chart: Chọn biểu đồ này nếu muốn thêm 1 chiều dữ liệu làm kích thước vào trực quan hoá. Biểu đồ bubble chart làm nổi bật trọng số của một số dữ liệu
* Tree map: Các loại chart (pie/ donut/ bar chart) chỉ hiệu quả khi biểu diễn cho dữ liệu nhỏ (<= 10 thành phần) -> tree map thay thế hiệu quả -> Có nhiều đối tượng và muốn xem những đối tượng nào đang chiếm tỷ trọng cao.

Hình ảnh hoá cho dữ liệu phân cấp, so sánh tỷ lệ giữa các danh mục thông qua kích thước của chúng. Sự tương quan giữa màu sắc và kích thước trong cấu trúc cây, người dùng có thể đọc được nhiều ý nghĩa

5 đặc tính quan trọng của Biểu đồ cây:

+ Tuỳ chỉnh tự động

+ Dải mầu phân cấp

+ Chú giải

+ Phân cấp sâu

+ Thao tác chọn nhanh

* Sandkey chart: Sử dụng để mô tả các luồng dịch chuyển từ 1 tệp giá trị (source) đến các tệp giá trị khác (Destination). Chiều rộng của luồng tỷ lệ theo một số lượng tính xác định (weight)
* Nắm bắt dòng chảy từ các nguồn, điểm đến và các bước trung gian trong quá trình di chuyển

1. Quy trình xử lý và làm sạch dữ liệu:

* Thu thập dữ liệu: Thu thập raw data từ các nguồn được xác định rõ ràng và có độ chính xác cao.
* Khai phá dữ liệu (EDA):

+ Thực hiện phân tích sơ bộ để hiểu cấu trúc của dữ liệu và phát hiện các vấn đề tiềm ẩn

Tổng quan về dữ liệu: Xem xét một số bản ghi đầu tiên và cuối cùng trong dữ liệu

Thống kê mô tả: Trung bình, trung vị, max, min, mod, std, …

Trực quan hoá để hình dung dữ liệu: Sử dụng biểu đồ cột, biểu đồ đường, boxplot, …

Vẽ phân phối dữ liệu

Tương quan -> xác định mối quan hệ giữa các biến thông qua biểu đồ tương quan giúp nhận biết xu hướng và quy luật (heatmap, scatter plot)

Thăm dò các biến quan trọng: Tập trung vào các biến quan trọng có thể ảnh hưởng đến mục tiêu của bạn và hiểu rõ hơn về chúng

Kiểm tra các giả định về phân phối, … của dữ liệu

+ Một số hàm trong python sử dụng để EDA: head(), describe(), info(), isnull(), sum(), count(), corrcoef(), hist(), scatterplot(), boxplot(), heatmap(), …

* Xác định và xử lý Missing value

+ Phân tích và xác định các giá trị thiếu trong dữ liệu

+ Xử lý giá trị thiếu bằng cách điền giá trị hoặc loại bỏ các bản ghi có giá trị thiếu

* Xử lý missing value:

+ Loại bỏ: Không khuyến khích cách làm này, chỉ thực hiện khi lượng dữ liệu bị khuyết dưới 3% tổng số quan sát trong một biến nhất định

+ Thay thế bằng giá trị khác:

TH1- Biến numerical: Có thể thay thế bằng các giá trị: 0, median, mean, mod, …

TH2- Biến categorical: Có thể nhóm vào 1 nhóm, đặt tên là missing

+ Điền giá trị bằng trung bình hoặc trung vị của nhóm (nhóm được xác định bởi 1 biến khác)

+ Sử dụng mô hình dự đoán học máy hoặc hồi quy để dự đoán giá trị bị thiếu dựa trên các biến khác trong dữ liệu

+ Kỹ thuật hot- deck imputation: Chọn một quan sát có giá trị tương tự và sử dụng giá trị của quan sát đó để điền vào giá trị thiếu.

+ Sử dụng giá trị trung bình trượt: Điền giá trị thiếu bằng giá trị trung bình của một số quan sát gần nhất trong dãy thời gian (với dữ liệu chuỗi thời gian)

+ Dùng giá trị gần kề: Sử dụng giá trị của quan sát gần kề nhất để điền vào giá trị thiếu

* Xử lý dữ liệu trùng lặp: Kiểm tra và loại bỏ các bản ghi trùng lặp trong dữ liệu
* Xử lý outlier:

+ Phân tích và xác định ngoại lệ trong dữ liệu

+ Quyết định liệu có nên loại bỏ ngoại lệ hay không, và nếu cần, thực hiện các biện pháp xử lý như thay thế giá trị ngoại lệ hoặc loại bỏ chúng hoặc tách chúng ra phân tích riêng

* Chuẩn hoá dữ liệu: Chuẩn hoá đơn vị đo lường, định dạng, hoặc thang đo để đảm bảo tính nhất quán và dễ dàng so sánh.

1. Các cách phát hiện outlier và xử lý

* Với dữ liệu numerical
* Với dữ liệu categorical
* Một số thuật toán phát hiện outlier: Kmean, HDBSCAN, Isolation forest, Local Outlier Factor, SVM one class

1. Phân tích hồi quy
2. Phân tích tương quan
3. Phân tích phân cụm
4. Các phương pháp kiểm định mô hình
5. Kiểm định giả thuyết thống kê

* Mục tiêu:
* Triển khai chiến dịch happycall cho thuê bao di động trả sau

(Happy call là gọi tới khách hàng để đánh giá về mức độ hài lòng của khách hàng về dịch vụ/ sản phẩm)

* Triển khai chiến dịch happycall cho thuê bao cố định
* Tăng thuê bao mua xổ số trên MyViettel
* Tăng thuê bao thẻ game GARENA trên MyViettel
* Tăng thuê bao di động chuyển sang trả sau theo đề xuất của mô hình phân tích dữ liệu
* Phát triển mới thuê bao FTTH (Internet cáp quang) theo đồ xuất của mô hình phân tích dữ liệu
* Tăng thuê bao FTTH chuyển đổi sang homewifi theo đề xuất của mô hình phân tích dữ liệu
* Giới thiệu:

Em chào anh chị, em là Lê Thị Ngọc Khánh, em tốt cử nhân ngành hệ thống thông tin quản lý của Đại Học Bách Khoa HÀ Nội. Hiện nay em đang làm OS Data Analyst tại BU ứng dụng và phân tích dữ liệu dịch vụ mới

Trong quá trình làm việc các kinh nghiệm mà em tích luỹ được là

* Làm việc với mô hình làm việc agile trong phát triển phần mềm
* Biết cách xây dựng KPI đáp ứng được chức năng, mục tiêu đặt ra, và lựa chọn loại biểu đồ để trực quan hoá số liệu của KPI đó. Cụ thể em đã xây dựng một số KPI của TV360 để hiển thị trên màn chi tiết đối tượng của hệ thống CCAI
* Xây dựng một số luồng tổng hợp dữ liệu cụ thể của màn chi tiết sự kiện và màn chi tiết đối tượng
* Cải thiện kỹ năng SQL, học được cách tối ưu hoá một luồng tổng hợp dữ liệu

+ Tối ưu hoá truy cập cơ sở dữ liệu: Chỉ trích xuất những dữ liệu cần thết

+ Lưu trữ các kết quả trước đó tránh việc tính toán lại các giá trị đã được tính toán trước đó

+ Giải phóng bộ nhớ khi không còn cần thiết

+ Chia nhỏ công việc thành các đơn vị nhỏ

+ Thiết lập một lịch trình cập nhật phù hợp với yêu cầu và tuần suất cập nhật của dữ liệu

* Hiểu về cách xử lý dữ liệu lớn sử dụng các công cụ như Hadoop, Spark
* Học được quy trình nghiệm thu các chức năng, thay đổi hàng tháng của hệ thống CCAI gồm:

+ Kế hoạch nghiệm thu: Ngày giờ, địa điểm, các biên bản và tài liệu liên quan

+ Đảm bảo đủ tài nguyên cho quá trình nghiệm thu bao gồm người kiểm thử, dữ liệu kiểm thử và môi trường kiểm thử

+ Kiểm tra các chức năng và thay đổi được triển khai trong hệ thống -> đảm bảo hoạt động đúng những yêu cầu không gây ra lỗi

+ Ghi lại và báo cáo mọi lỗi hoặc vấn đề phát sinh trong quá trình nghiệm thu

+ Sau khi tất cả các vấn đề được giải quyết -> triển khi trên môi trường thật

* Đầu ra cho quá trình này là cần đảm bảo hệ thống CCAI hoạt động, hiển thị đúng với từng chức năng như:

+ Màn Dashboard:

+ Màn Sự kiện:

+ Màn Insight Dashboard:

+ Màn Dự đoán:

* Học được quy trình nghiệm thu các chức năng, thay đổi hàng tháng của hệ thống CCAI gồm:

+ Kế hoạch nghiệm thu: Ngày giờ, địa điểm, các biên bản và tài liệu liên quan

+ Đảm bảo đủ tài nguyên cho quá trình nghiệm thu bao gồm người kiểm thử, dữ liệu kiểm thử và môi trường kiểm thử

+ Kiểm tra các chức năng và thay đổi được triển khai trong hệ thống -> đảm bảo hoạt động đúng những yêu cầu không gây ra lỗi

+ Ghi lại và báo cáo mọi lỗi hoặc vấn đề phát sinh trong quá trình nghiệm thu

+ Sau khi tất cả các vấn đề được giải quyết -> triển khi trên môi trường thật

* Đầu ra cho quá trình này là cần đảm bảo hệ thống CCAI hoạt động, hiển thị đúng với từng chức năng:

+ Màn Dashboard:

+ Màn Sự kiện:

+ Màn Insight Dashboard:

+ Màn Dự đoán:

Mỗi dashboard mình sẽ hiển thị với 1 user thuộc 1 domain là các giá trị khác nhau và để theo dõi chỉ số kinh doanh của các domain dịch vụ đấy

Domain bussiness:

- TV360 khác với truyền thông nhắn tin, khác với truyền thông bán các dịch vụ của chuyển dịch số

Sự kiện:

Mục tiêu- link mocha

Sự kiện các với sự kiện chuyển đổi ở chỗ một sự kiện diễn ra là mình theo dõi khách hàng theo sự kiện đấy

Sự kiện chuyển đổi là hai sự kiện bắt nhau

(Giống như theo dõi hành trình:

Sự kiện sẽ là đường đi, các điểm đi sẽ là sự kiện chuyển đổi

Theo dòng là sự kiện

Khi đạt đến một ngưỡng thì sẽ là sự kiện chuyển đổi

Mục đích của phần này:

Ví dụ:

Khách hàng mỗi tháng tiêu 10K -> mua 2 gói -> nhắn 2 tin

Nếu như đủ hết 3 sự kiện trên thì mình gọi callbot

Nếu đủ 4 điều kiện này mà khách hàng không reject -> chuyển tư vấn viên

Đấy Tác dụng của sự kiên cộng với segmentation

Sắp tới em mong muốn bổ sung vào CCAI là chức năng hình thành luồng tương tác dữ liệu tổng thể để theo dấu vòng đời khách hàng

* Phục vụ chủ yếu là phân tích và đánh giá để đưa ra các phạm vi chiến lược về kinh doanh, xây dựng hoạch định kinh doanh và phân tích phạm vi tương tác của dữ liệu

Ngoài ra trong quá trình học tập và làm dự án cá nhân tại trường em còn tích luỹ cho mình một số kỹ năng, kinh nghiệm như:

* Thu thập, xử lý dữ liệu
* Các cách phát hiện outlier và xử lý
* Một số phân tích như: Phân tích thống kê, phân tích mô tả, phân tích hồi quy, phân tích tương quan, phân tích phân cụm
* Kiểm định mô hình
* Kiểm định giả thuyết thống kê
* Mô hình phân tích đa chiều OLAP