**Nguyễn Tử Trung - 21020415**

**Chương 5:**

**Smart Analytics, Machine Learning, and AI on Google Cloud**

**Phần 1: Giới thiệu về phân tích và AI**

* AI là gì?
* Định nghĩa AI, ML, và Học Sâu:
* AI được mô tả như một lĩnh vực tập trung vào việc tạo ra máy móc có khả năng tự động hành động và suy nghĩ giống như con người.
* ML là cách sử dụng thuật toán để rút ra thông tin dự đoán từ dữ liệu, hỗ trợ quyết định hướng tương lai.
* Học sâu là một dạng của ML hoạt động tốt với dữ liệu không cấu trúc như hình ảnh, giọng nói, và văn bản tự nhiên.
* Sự Khác Biệt Giữa BI và ML:
* Trong BI (Business Intelligence), thông tin thường dựa trên dữ liệu lịch sử để tạo ra báo cáo và bảng điều khiển.
* ML tập trung vào việc tạo ra thông tin dự đoán có thể mở rộng và lặp lại, hỗ trợ quyết định trong tương lai.
* Ứng Dụng Cụ Thể của ML:
* Keller Williams sử dụng AutoML Vision để nhận diện đặc điểm của căn nhà trong ảnh quảng cáo, giúp môi giới bất động sản nhanh chóng đăng ký nhà và khách hàng dễ dàng tìm kiếm theo yêu cầu.
* Kewpie sử dụng học máy để phát hiện chất lượng kém của viên khoai tây trong thức ăn cho trẻ em, giúp giảm công việc kiểm tra thủ công.
* AutoML Vision và Học Máy không cần mã lập trình:
* AutoML Vision được sử dụng mà không cần viết mã, đơn giản hóa quá trình tạo mô hình dự đoán từ dữ liệu hình ảnh.
* Thực tế, việc tạo mô hình có thể được thực hiện thông qua việc đưa hình ảnh và video sẵn có vào mô hình AutoML Vision mà không cần một dòng mã lập trình nào.
* Thách Thức và Lợi Ích của ML trong Doanh Nghiệp:
* Thách thức: Khả năng của con người giới hạn trong việc đưa ra các quyết định dự đoán cho hàng loạt sản phẩm và khu vực khác nhau.
* Lợi ích: Sử dụng ML để tự động hóa việc đưa ra thông tin dự đoán từ dữ liệu, tạo ra quyết định có thể lặp lại và mở rộng.
* Từ phân tích dữ liệu đến đưa ra quyết định
* Mục tiêu của bài học là khám phá cách Trí tuệ Nhân tạo (AI) có thể giúp chuyển đổi quá trình phân tích dữ liệu tạm thời thành quyết định dựa trên dữ liệu, hỗ trợ quyết định hơn và hiệu quả hơn.
* Một ví dụ minh họa được đưa ra với chủ doanh nghiệp cho thuê xe đạp tại London để làm rõ quá trình từ phân tích tạm thời đến quyết định dựa trên dữ liệu. Chủ doanh nghiệp có thể sử dụng phân tích dữ liệu để xem xét xu hướng trong dữ liệu lịch sử, bao gồm thời lượng thuê trung bình hàng ngày theo thời gian và các yếu tố khác như trạm thuê và khoảng thời gian cụ thể.
* Mô hình Học Máy có khả năng áp dụng các thuật toán khác nhau để tạo ra thông tin với quy mô và tốc độ lớn hơn so với sự can thiệp của con người, giúp giải quyết vấn đề của chủ doanh nghiệp.
* Ví dụ minh họa trong bài giảng sử dụng một mô hình hồi quy tuyến tính để dự đoán nhu cầu thuê xe đạp theo giờ.
* Một số mô hình ML
* Mô Hình Tùy Chỉnh thông qua TensorFlow và Vertex AI:
* Người sử dụng có trình độ cao và muốn kiểm soát nhiều hơn về việc xây dựng và huấn luyện mô hình Học Máy của mình có thể sử dụng các công cụ cung cấp mức độ linh hoạt mà họ đang tìm kiếm.
* Cách tiếp cận này liên quan đến việc phát triển các mô hình tùy chỉnh thông qua một thư viện Học Máy như TensorFlow, được hỗ trợ trên Vertex AI.
* Sử Dụng AutoML:
* Google cung cấp tùy chọn AutoML cho những người dùng không có kiến thức sâu về học máy.
* AutoML cho phép xây dựng các mô hình học máy tùy chỉnh phù hợp với nhu cầu kinh doanh cụ thể mà không cần viết mã TensorFlow.
* Người sử dụng có thể tích hợp những mô hình này vào ứng dụng và trang web mà không cần chạy bất kỳ dòng mã TensorFlow nào.
* Mô Hình Học Máy Đã Đào Tạo Trước (Pretrained Models):
* Google cung cấp một loạt các mô hình đã đào tạo trước, nghĩa là không cần mang theo dữ liệu của riêng bạn.
* Những mô hình này đã sẵn sàng để sử dụng ngay lập tức trong ứng dụng theo các cách mà các API tương ứng được thiết kế để hỗ trợ.
* Các mô hình đã đào tạo trước là cách tốt để thay thế đầu vào từ người dùng bằng học máy.

**Phần 2: Chuẩn bị API của mô hình ML cho dữ liệu phi cấu trúc**

* Khó khăn khi làm việc với dữ liệu phi cấu trúc
* Dữ liệu Phi Cấu Trúc là Gì:
* Dữ liệu phi cấu trúc là dữ liệu không theo mô hình cấu trúc truyền thống, chẳng hạn như các bảng cơ sở dữ liệu quan hệ. Thay vào đó, đó có thể là các dạng dữ liệu như văn bản, hình ảnh, âm thanh, video, không tuân theo cấu trúc rõ ràng.
* Khó Khăn khi Làm Việc với Dữ Liệu Phi Cấu Trúc:
* Việc làm việc với dữ liệu phi cấu trúc gặp nhiều khó khăn do thông tin trong dữ liệu thường phức tạp và đa dạng.
* Trong ví dụ của hình ảnh, việc trích xuất thông tin như ngôn ngữ, nội dung bài viết, năm xuất bản, thể thao, cờ và các yếu tố khác là thách thức.
* Dữ liệu phi cấu trúc thường đòi hỏi nhiều công sức khi phải xử lý và phân tích, đặc biệt là khi làm việc với ảnh y khoa có độ phân giải cao.
* Các mô hình học máy được đào tạo trước có thể hữu ích, nhưng kết quả có thể không tốt nếu dữ liệu của bạn không nằm trong phạm vi của mô hình đào tạo trước đó
* Một vài mô hình ML để làm việc với dữ liệu phi cấu trúc
* Cloud Natural Language API của Google để xử lý dữ liệu phi cấu trúc dưới dạng văn bản.
* Chức Năng của Cloud Natural Language API:
* API cung cấp nhiều tính năng cho việc thực hiện phân tích văn bản, bao gồm:
* Phân Tích Cú Pháp (Syntactic Analysis): Phân tích cú pháp phân chia văn bản thành các "tokens" như từ và câu, cung cấp thông tin về cấu trúc nội tại và vai trò của token trong câu.
* Phân Tích Thực Thể (Entity Analysis): Nhận diện người, địa điểm, tổ chức, sự kiện, tác phẩm, sản phẩm tiêu dùng, số điện thoại, địa chỉ, ngày tháng, và số trong văn bản.
* Phân Tích Tâm Trạng (Sentiment Analysis): Xác định tâm trạng của tác giả và biểu diễn dưới dạng điểm số và độ lớn.
* Phân Tích Tâm Trạng của Thực Thể (Entity Sentiment Analysis): Kết hợp cả phân tích thực thể và phân tích tâm trạng để xác định tâm trạng đối với mỗi thực thể trong văn bản.

**Phần 3: Phân tích Big Data với các notebook**

* Notebook là gì
* Notebooks là môi trường phát triển chuyên dụng, giúp việc viết mã cho phân tích dữ liệu và machine learning trở nên hiệu quả hơn.
* Chúng tích hợp một cách mượt mà giữa bình luận, biểu đồ và mã nguồn.
* Lợi ích của Notebooks:
* Thay vì sử dụng các công cụ phát triển phần mềm tiêu chuẩn, notebooks cung cấp môi trường linh hoạt cho việc thực hiện phân tích dữ liệu và lập trình machine learning.
* Notebooks tổ chức mã nguồn thành các ô (cells) có thể thực thi độc lập, giúp thực hiện các thay đổi nhỏ một cách linh hoạt và hiệu quả.
* Chúng hỗ trợ tính năng ghi chú, biểu đồ và mã nguồn trong một tài liệu duy nhất.
* Môi Trường Làm Việc:
* Notebooks cung cấp môi trường làm việc giống như việc sử dụng Google Docs, nơi bạn có thể tương tác trực tiếp với mã nguồn mà không cần phải tạo ra các file tài liệu riêng biệt.
* Cho phép việc hợp tác dễ dàng, giúp tránh được việc gửi đi nhiều bản sao và tiết kiệm thời gian so với quá trình làm việc truyền thống.
* Sử Dụng Trong Đám Mây (Cloud):
* Notebooks có thể chạy trong môi trường đám mây, giúp dễ dàng phát triển chung và truy cập ngay cả khi máy tính cá nhân không hoạt động.
* Cung cấp khả năng chia sẻ notebook thông qua URL hoặc sử dụng các hệ thống kiểm soát phiên bản như Git.
* Kết Hợp với Công Nghệ Máy Học:
* Sử dụng Python notebooks, đặc biệt là Jupyter Notebooks, được tích hợp với các thư viện máy học chuẩn.
* Hỗ trợ các tùy chọn cấu hình phần cứng, bao gồm việc thêm GPU vào các máy ảo để tăng tốc các công việc như machine learning và xử lý dữ liệu.
* Tích Hợp với BigQuery:
* Notebooks cung cấp kết nối với BigQuery, giúp dễ dàng truy xuất dữ liệu từ BigQuery và tích hợp vào notebook cho việc phân tích.
* Trực tiếp truy xuất dữ liệu từ BigQuery vào notebook
* Magic Functions trong Jupyter Labs:
* Magic functions là các hàm đặc biệt trong Jupyter Labs cho phép thực thi các lệnh hệ thống từ các ô notebook.
* Các magic functions có thể được sử dụng để kiểm tra nội dung thư mục hiện tại và định nghĩa các magic function tùy chỉnh.
* BigQuery Magic Function:
* Trong bài giảng, giới thiệu về BigQuery magic function, một magic function đặc biệt cho việc thực hiện các truy vấn BigQuery trực tiếp từ notebook cells.
* Magic function này hữu ích để kiểm tra định dạng truy vấn, tính đúng đắn và xem kết quả.
* Lưu Kết Quả Truy Vấn vào Pandas DataFrame:
* BigQuery magic function cho phép lưu kết quả của truy vấn vào một Pandas DataFrame để bạn có thể tiếp tục xử lý dữ liệu.
* Pandas là một thư viện số học cho Python, hiển thị dữ liệu trong cấu trúc bảng và cung cấp thông tin về metadata.
* Xử Lý Dữ Liệu Trong Pandas:
* Pandas DataFrame giúp hiển thị dữ liệu dưới dạng bảng, cung cấp thông tin về các loại đối tượng và tạo thống kê tóm tắt về các cột.
* Việc này giúp dễ dàng thực hiện các thao tác phức tạp như mô tả dữ liệu, xác định loại đối tượng của các cột và sinh thống kê tóm tắt.
* Hạn Chế Về Bộ Nhớ:
* Lưu ý rằng bạn đang làm việc trong một notebook và có giới hạn về bộ nhớ.
* Khi truy xuất dữ liệu từ BigQuery, hạn chế kích thước bằng cách sử dụng kỹ thuật lấy mẫu để truy xuất một phần nhỏ của dữ liệu.
* Ví Dụ về Thực Hiện Truy Vấn và Tạo Biểu Đồ:
* Một ví dụ được thể hiện, trong đó một truy vấn BigQuery được thực hiện để truy xuất dữ liệu và tạo một biểu đồ cột đơn giản từ Pandas DataFrame.

**Phần 4: Production ML Pipelines**

* Cách để tạo Machine Learning trên Google Clouds
* Ba cách tiếp cận chính trong ML trên GC: sử dụng các mô hình được đào tạo trước, xây dựng các mô hình tùy chỉnh và triển khai chúng bằng Vertex AI
* Vertex AI là một dịch vụ được quản lý hoàn toàn dành cho các mô hình học máy tùy chỉnh, xử lý cả dự đoán đào tạo và cung cấp. Nó cung cấp khả năng mở rộng từ thử nghiệm đến sản xuất và tích hợp với TensorFlow để chuyển đổi dữ liệu đầu vào và điều chỉnh siêu tham số.
* Các mô hình được phát triển bằng Vertex AI có thể được triển khai cho các dịch vụ dự đoán tự động điều chỉnh quy mô dựa trên nhu cầu của khách hàng. Nó đơn giản hóa quy trình, cho phép các nhà khoa học dữ liệu đào tạo và triển khai các mô hình sản xuất trực tiếp từ sổ ghi chép bằng một số lệnh.
* Đường ống Vertex AI
* Xây dựng đường ống dữ liệu:
* Kỹ sư dữ liệu là chuyên gia trong việc thiết kế và xây dựng đường ống dữ liệu.
* Quy trình ML yêu cầu nhiều bước xử lý dữ liệu khác nhau, bao gồm làm sạch, chuyển đổi dữ liệu và kỹ thuật tính năng. Kỹ sư dữ liệu đảm bảo các bước này được tích hợp liền mạch vào quy trình ML.
* Hỗ trợ đường ống ML:
* Quy trình ML bao gồm các quy trình công việc phức tạp, gồm nhiều bước và Kỹ sư dữ liệu đóng góp vào việc thiết kế và triển khai các quy trình này.
* Họ tạo ra các quy trình làm việc bao gồm các nhiệm vụ như tải tập dữ liệu, làm sạch dữ liệu, tạo tính năng, đào tạo mô hình và đánh giá hiệu suất của mô hình.
* Đảm bảo khả năng tái tạo và khả năng mở rộng:
* Kỹ sư dữ liệu đảm bảo rằng các đường dẫn ML được thiết kế để có thể tái tạo, kiểm tra được, tiết kiệm chi phí và có thể mở rộng.
* Họ quản lý cơ sở hạ tầng và kiến trúc của quy trình để xử lý các khối lượng công việc khác nhau một cách hiệu quả.
* Quản lý các bước quy trình làm việc:
* Quy trình mô hình hóa quy trình làm việc dưới dạng tập hợp các bước được kết nối và Kỹ sư dữ liệu đóng vai trò quản lý từng bước.
* Họ đảm bảo rằng mỗi bước lấy đầu vào từ các bước trước đó, thực hiện các phép tính cần thiết và tạo đầu ra cho các thành phần tiếp theo.
* Tối ưu hóa và giám sát:
* Kỹ sư dữ liệu đóng vai trò quan trọng trong việc tối ưu hóa hiệu suất đường ống và giám sát hiệu quả của nó.
* Chúng cung cấp các điểm kiểm soát để giám sát và tối ưu hóa liên tục tài sản ML sản xuất.
* AI Hub
* AI của Vertex có thể được chia sẻ và phân phối thông qua AI Hub, đóng vai trò như một kho lưu trữ tập trung cho các thành phần máy học. AI Hub cho phép người dùng truy cập và triển khai các tài sản ML do người khác tạo, thúc đẩy sự hợp tác, hiệu quả và tránh sự dư thừa trong việc xây dựng các thành phần đã tồn tại.
* Kho lưu trữ các thành phần ML:
* AI Hub là kho lưu trữ tập trung nơi người dùng có thể khám phá, chia sẻ và triển khai các thành phần máy học.
* Tránh dư thừa:
* Người dùng có thể tránh phải phát minh lại bánh xe bằng cách tìm và sử dụng các thành phần mà người khác đã xây dựng và tối ưu hóa.
* Hỗ trợ cho các loại tài sản khác nhau:
* AI Hub lưu trữ nhiều loại nội dung khác nhau, bao gồm toàn bộ quy trình Kubeflow, sổ ghi chép Jupyter, mô-đun TensorFlow, các mô hình, dịch vụ và hình ảnh VM được đào tạo đầy đủ.
* Thông tin tài sản điển hình:
* Mỗi nội dung trên AI Hub cung cấp thông tin về quy trình, bao gồm thông tin chi tiết về đầu vào và đầu ra, cùng với các tùy chọn để tải xuống.
* Phạm vi công khai và hạn chế:
* AI Hub phân loại tài sản thành hai phạm vi: tài sản công và tài sản bị hạn chế.
* Tài sản công cộng có sẵn cho tất cả người dùng AI Hub, thúc đẩy sự chia sẻ và cộng tác rộng rãi.
* Nội dung trong phạm vi bị hạn chế bao gồm các thành phần AI mà người dùng đã tải lên và những thành phần được chia sẻ trong các tổ chức hoặc nhóm cụ thể. Điều này cho phép truy cập có kiểm soát vào tài sản.​

**Phần 5: Tùy chỉnh mô hình xây dựng trên SQL và BigQuery ML**

* BigQuery ML cho Quick Model Building
* Tạo mô hình tùy chỉnh:
* Không giống như API ML, BigQuery ML cho phép tạo các mô hình tùy chỉnh.
* Tổng quan về AutoML:
* AutoML được đề cập ngắn gọn như một công cụ tận dụng các mô hình dựng sẵn của Google trong một số trường hợp nhất định. Nó cho phép người dùng xây dựng các mô hình tùy chỉnh từ đầu bằng cách sử dụng phương pháp học chuyển giao và một hình thức tìm kiếm kiến trúc thần kinh.
* Quy trình làm việc BigQuery ML: Quy trình làm việc trong BigQuery ML bao gồm một số bước từ đầu đến suy luận:
* Viết truy vấn trên dữ liệu được lưu trữ trong BigQuery để trích xuất dữ liệu đào tạo.
* Tạo một mô hình chỉ định loại mô hình và siêu tham số.
* Huấn luyện và đánh giá mô hình.
* Đưa ra dự đoán bằng cách sử dụng mô hình đã đào tạo về dữ liệu được trích xuất từ BigQuery.
* Ví dụ về trường hợp sử dụng - Tin tức về hacker:
* Một ví dụ được trình bày trong đó mục tiêu là xác định nguồn xuất bản (Techcrunch, GitHub hoặc The New York Times) của các bài viết sử dụng BigQuery ML.
* Một truy vấn đặc biệt được viết để khám phá dữ liệu, trích xuất thông tin từ tiêu đề và URL bằng cách sử dụng các hàm biểu thức chính quy.
* Cú pháp tạo mô hình:
* Cú pháp để tạo một mô hình bao gồm việc sử dụng câu lệnh CREATE OR REPLACE MODEL, chỉ định tên mô hình, các tùy chọn và siêu tham số.
* Truy vấn xác định tập huấn luyện tuân theo mệnh đề AS.
* Số liệu đánh giá mô hình:
* Các số liệu như độ chính xác, thu hồi, độ chính xác, f1\_score, log\_loss và roc\_curve đều được đề cập.
* Các số liệu đánh giá là cần thiết để đánh giá hiệu suất của mô hình.
* Ma trận hỗn loạn:
* Ma trận nhầm lẫn được giới thiệu như một công cụ để trực quan hóa nơi mô hình đưa ra dự đoán không chính xác.
* Triển khai mô hình:
* Sau khi mô hình đáp ứng các yêu cầu, mô hình sẽ tự động có sẵn để phân phát các dự đoán trong BigQuery.
* Người dùng không cần lo lắng về việc triển khai mô hình riêng biệt; nó đã sẵn sàng để dự đoán trong BigQuery.
* Ví dụ dự đoán hàng loạt:
* Một ví dụ về dự đoán hàng loạt sử dụng BigQuery ML được cung cấp, cho thấy cách mô hình dự đoán nguồn của một bài viết dựa trên tiêu đề của nó.
* Mô hình hỗ trợ
* Phân loại tuyến tính (Hồi quy logistic):
* Ví dụ: Dự đoán chuyến bay đến (kết quả nhị phân - đúng giờ hay không).
* Tùy chọn: Chính quy hóa cho người dùng nâng cao.
* Mạng thần kinh sâu (DNN):
* Cho phép mô hình hóa tốt hơn các mối quan hệ phi tuyến trong dữ liệu
* Ví dụ: Mô hình khấu hao ô tô.
* Triển khai: Chỉ định `DNN` làm loại mô hình trong tiêu đề `TẠO HOẶC THAY THẾ MODEL`.
* Cây được tăng cường độ dốc (XGBoost):
* Ví dụ: Chỉ định `bộ phân loại cây tăng cường` cho các tác vụ phân loại.
* Siêu tham số: Bao gồm các tùy chọn như độ sâu cây tối đa.
* Khái niệm: Lặp lại việc khớp cây quyết định với phần dư của cây quyết định trước đó.
* Mô hình hồi quy:
* Hồi quy tuyến tính: Dự đoán giá vé taxi dựa trên các đặc điểm như giờ trong ngày và địa điểm.
* Bộ hồi quy DNN: Chỉ định `bộ hồi quy DNN` và xác định các đơn vị ẩn cho các mô hình hồi quy phức tạp hơn.
* XGBoost: Chỉ định `bộ hồi quy cây được tăng cường` cho các tác vụ hồi quy.
* Nhập mô hình TensorFlow:
* Khả năng nhập các mô hình TensorFlow được đào tạo trước vào BigQuery để dự đoán hàng loạt.
* Ví dụ: Đang tải mô hình, chỉ định tên mô hình, đường dẫn mô hình trong Cloud Storage và đặt loại mô hình là TensorFlow.
* Dự đoán: Sử dụng câu lệnh `ML.PREDICT` để tạo dự đoán.
* Mô hình khuyến nghị (Hệ số ma trận):
* Ví dụ: Xây dựng công cụ đề xuất trong BigQuery ML.
* Loại mô hình: Phân tích nhân tử ma trận.
* Khái niệm: Phân tách ma trận tương tác giữa người dùng và mục thành các ma trận hình chữ nhật có chiều thấp hơn.
* Dự đoán: Sử dụng câu lệnh `ML.PREDICT` để đưa ra các đề xuất.
* Học không giám sát (Phân cụm K-means):
* Khám phá các mẫu trong dữ liệu mà không cần dự đoán nhãn cụ thể.
* Ví dụ: Phân cụm dữ liệu thuê xe đạp thành 4 cụm.
* Tiền xử lý: Chuẩn hóa tính năng, loại bỏ trường ID.
* Truy vấn cụm: Sử dụng truy vấn SQL để xác định phân cụm (CENTROID\_ID).
* Phân tích: Khám phá phạm vi tính năng và số liệu thống kê tổng hợp để hiểu các đặc điểm của cụm.
* Bảng cheat:
* Cung cấp bảng tóm tắt nhanh với cú pháp đào tạo mô hình, đánh giá mô hình, đưa ra dự đoán và các nhiệm vụ cần thiết khác trong BigQuery ML.

**Phần 6: Tùy chỉnh mô hình xây dựng trên AutoML**

* AutoML là gì?
* Vị trí của AutoML trong các Sản phẩm ML của Google Cloud:
* AutoML nằm giữa các mô hình được tùy chỉnh hoàn toàn (Vertex AI, BigQuery ML) và các mô hình được đào tạo trước (API Vision, API chuyển giọng nói thành văn bản).
* Cung cấp sự cân bằng giữa việc tùy chỉnh mô hình và tính dễ sử dụng mà không yêu cầu chuyên môn về mã hóa.
* Những thách thức của mô hình tùy chỉnh:
* Các mô hình tùy chỉnh (Vertex AI, BigQuery ML) yêu cầu kiến thức chuyên môn về máy học và kinh nghiệm viết mã.
* Việc đào tạo những mô hình này có thể tốn thời gian.
* Lợi ích của AutoML:
* Đào tạo một mô hình cụ thể cho dữ liệu người dùng mà không yêu cầu mã.
* Thu hẹp khoảng cách giữa các mô hình hoàn toàn tùy chỉnh và các mô hình được đào tạo trước.
* Quy trình làm việc AutoML:
* Chia thành ba giai đoạn: Đào tạo, Triển khai và Phục vụ.
* Giai đoạn đào tạo bao gồm việc chuẩn bị một tập dữ liệu, phân tích và xác nhận nó, huấn luyện mô hình và đánh giá tính hiệu quả của nó với dữ liệu thử nghiệm.
* Giai đoạn triển khai bao gồm quản lý các mô hình, loại bỏ những mô hình không sử dụng và lưu trữ mô hình để dự đoán.
* AutoML đơn giản hóa các giai đoạn triển khai và phục vụ so với ML truyền thống.
* Chuẩn bị dữ liệu cho AutoML:
* Dữ liệu có thể được chuẩn bị trực tiếp trong giao diện người dùng web hoặc được tập hợp trong tệp CSV.
* Định dạng tệp CSV có nhãn, nhận dạng tệp nguồn và gán dữ liệu tùy chọn vào các nhóm đào tạo, xác thực hoặc kiểm tra.
* AutoML kiểm tra và phân tích tập dữ liệu đã chuẩn bị sẵn sàng.
* Đào tạo một mô hình tùy chỉnh:
* AutoML sử dụng tập dữ liệu đã chuẩn bị sẵn để huấn luyện mô hình tùy chỉnh thông qua học tập có giám sát.
* Quá trình đào tạo bao gồm nhiều giai đoạn, sửa lỗi và giảm thiểu tổng số lỗi để tạo ra một mô hình hiệu quả.
* Đánh giá mô hình:
* Sử dụng dữ liệu của nhóm thử nghiệm để đánh giá hiệu suất của mô hình.
* Cung cấp báo cáo đánh giá với các chỉ số cụ thể cho loại mô hình.
* Kích hoạt và vòng đời của các mẫu tùy chỉnh:
* Các mô hình sẽ được kích hoạt tự động và có thể cần thời gian khởi động.
* Các mô hình tùy chỉnh chỉ là tạm thời và cuối cùng sẽ bị xóa.
* Các mô hình không được sử dụng để dự đoán sẽ tự động bị xóa sau một khoảng thời gian.
* Đưa ra dự đoán bằng các mô hình tùy chỉnh:
* Giao diện phân loại được cung cấp thông qua URI.
* Dự đoán có thể được thực hiện bằng cách sử dụng Giao diện người dùng Web, dòng lệnh (CURL) hoặc thư viện máy khách cho Python, Java và Node.js.
* Phản hồi JSON bao gồm tên hiển thị, phân loại và điểm tin cậy.
* Chiến lược ứng dụng:
* AutoML giảm bớt nỗ lực so với ML truyền thống trong việc tạo mô hình.
* Cho phép tạo các mô hình tùy chỉnh nhỏ hơn, chuyên biệt hơn.
* Nhiều mô hình chuyên biệt có thể được kết hợp theo chương trình để có kết quả hiệu quả hơn.
* Kết luận và chiến lược đề xuất:
* Bắt đầu với các dịch vụ AI dựng sẵn.
* Sử dụng AutoML cho các mô hình tùy chỉnh.
* Chia vấn đề thành các phần chuyên biệt và sử dụng nhiều mô hình tùy chỉnh.
* Sử dụng dịch vụ máy học và AI cho các tính năng nâng cao khi cần.
* AutoML Vision
* Tổng quan về Tầm nhìn AutoML:
* AutoML Vision là sản phẩm AutoML được thiết kế cho dữ liệu hình ảnh.
* Chuyên đào tạo mô hình phân loại ảnh.
* Tải và định dạng dữ liệu:
* Dữ liệu, bao gồm tệp CSV và tệp hình ảnh, có thể được tải từ bộ lưu trữ đám mây hoặc tải lên từ máy tính cục bộ bằng cách nhập.
* Các định dạng tệp hình ảnh được hỗ trợ để đào tạo bao gồm JPEG và PNG.
* Hình ảnh có thể có kích thước lên tới 30 megabyte.
* Hình ảnh cần được chuyển đổi sang mã hóa base64 để đào tạo.
* Nhận dạng và hạn chế tập tin:
* Dịch vụ nhận dạng các tệp JPEG, PNG và GIF lên tới 1,5 megabyte.
* Mô hình được đào tạo hoạt động tốt nhất khi số lượng hình ảnh cho nhãn phổ biến nhất nhiều hơn tối đa 100 lần so với nhãn ít phổ biến nhất.
* Khuyến nghị loại bỏ nhãn tần số rất thấp để có hiệu suất mô hình tốt hơn.
* Tùy chọn ghi nhãn:
* Hình ảnh có thể được gắn nhãn trong giao diện người dùng Web.
* Dịch vụ gắn nhãn do con người của Google có thể được sử dụng cho hơn 100 hình ảnh chưa được gắn nhãn.
* Đánh giá mức độ sẵn sàng của mô hình:
* AutoML Vision tạo ma trận nhầm lẫn cho tối đa 10 nhãn để đánh giá mức độ sẵn sàng của mô hình được đào tạo.
* Nếu có nhiều hơn 10 nhãn, ma trận sẽ bao gồm 10 nhãn có dự đoán sai nhất.
* Định dạng dữ liệu đào tạo:
* Các định dạng tệp đào tạo bao gồm JPEG, PNG, WebP, GIF, BMP, TIFF, ICO, tối đa 30 megabyte.
* Yêu cầu dịch vụ hỗ trợ các tệp JPEG, PNG hoặc GIF lên tới 1,5 megabyte.
* Chất lượng của tầm nhìn
* Chất lượng mô hình bị ảnh hưởng bởi việc lựa chọn dữ liệu huấn luyện.
* Đào tạo về các hình ảnh có đặc tính tương tự với hình ảnh dùng để phân loại (ví dụ: độ phân giải, ánh sáng, tiêu điểm và mức độ chi tiết tương tự).
* Điểm số có độ nhầm lẫn cao, độ chính xác trung bình thấp có thể cho thấy nhu cầu về dữ liệu đào tạo bổ sung hoặc nhãn không nhất quán.
* Phấn đấu đạt điểm tốt thay vì hoàn hảo để đảm bảo mô hình hoạt động tốt hơn dữ liệu kiểm tra.
* Điểm hoàn hảo hoặc có độ chính xác trung bình rất cao có thể chỉ ra các vấn đề và trong những trường hợp như vậy, hãy tăng tính đa dạng của hình ảnh trong tập dữ liệu đã chuẩn bị.
* AutoML Xử lý ngôn ngữ tự nhiên
* Tổng quan về ngôn ngữ tự nhiên AutoML:
* Chuyên đào tạo mô hình cho dữ liệu văn bản.
* Được sử dụng để phân loại văn bản, chẳng hạn như xác định xem bài viết có liên quan đến thể thao hay chính trị hay không.
* Nhận dạng và lưu trữ văn bản:
* Văn bản có thể nằm trong dòng trong tệp CSV hoặc được lưu trữ ở dạng tệp .txt hoặc định dạng nén và nén.
* Tệp CSV bao gồm đường dẫn lưu trữ đám mây tới tài liệu.
* Tài liệu phải là văn bản chuẩn và không hỗ trợ Unicode.
* Kích thước tài liệu có thể dao động từ một câu đến tối đa 128 kilobyte.
* Đánh giá nhãn và mẫu:
* Hỗ trợ từ hai đến 100 nhãn.
* Đánh giá mô hình dựa trên độ chính xác trung bình, dao động từ 0,5 đến 1,0 (giá trị cao hơn cho thấy phân loại chính xác hơn).
* Đường cong ngưỡng tin cậy được cung cấp để mô tả đặc điểm phân loại dương tính giả so với dương tính thực.
* Ma trận nhầm lẫn được bao gồm cho các mô hình áp dụng một nhãn cho mỗi tài liệu.
* Đánh giá cho mỗi nhãn có thể được xem.
* Vòng đời mô hình và đào tạo lại:
* Các mô hình không được sử dụng trong 60 ngày sẽ bị xóa; nếu được sử dụng, chúng sẽ bị xóa sau 6 tháng.
* Việc đặt trước một mô hình yêu cầu phải đào tạo lại do liên tục cải tiến và cập nhật về phương pháp đào tạo và phục vụ.
* Các thay đổi có thể không tương thích ngược.
* Cải thiện chất lượng mô hình:
* Độ nhầm lẫn cao và điểm chính xác trung bình thấp có thể cho thấy cần có dữ liệu bổ sung hoặc cách sử dụng nhãn không nhất quán.
* Các phương pháp cải thiện đánh giá bao gồm thêm nhiều tài liệu hơn cho các nhãn cụ thể, tăng tính đa dạng của tài liệu và loại bỏ các nhãn kém hữu ích hoặc chất lượng thấp.
* AutoML Tables
* Giới thiệu về Bảng AutoML:
* Được thiết kế cho dữ liệu dạng bảng có cấu trúc (điển hình của bảng tính).
* Được phát triển bởi Google Cloud và Nhóm Google Brain.
* Ứng dụng liên quan đến việc sử dụng kiến trúc ban đầu cho khả năng tìm kiếm và điều chỉnh nó cho các vấn đề dịch thuật và phân loại hình ảnh.
* Trường hợp sử dụng ví dụ: Thử thách đề xuất giá của Mercari:
* Áp dụng cho thử thách dự đoán giá sản phẩm của Mercari.
* Thể hiện thành tích ấn tượng dù chỉ sau 1 giờ luyện tập.
* Hiệu suất ổn định đạt được nhanh chóng, giúp tiết kiệm chi phí.
* Nhập dữ liệu và nguồn:
* Dữ liệu có thể được nhập qua BigQuery, mảng và cấu trúc hỗ trợ hoặc qua tệp CSV cục bộ hoặc trên Cloud Storage.
* Yêu cầu dữ liệu: 1.000 đến 100 triệu hàng, 2 đến 1.000 cột và kích thước từ 100 gigabyte trở xuống.
* Quy trình xây dựng mô hình:
* Chọn các tính năng và chỉ định cột mục tiêu.
* Giai đoạn xác thực dữ liệu kiểm tra các vấn đề như giá trị null, giá trị ngoại lệ và các cột không tương quan.
* Đào tạo và lập ngân sách:
* Việc đào tạo có thể được thiết lập trong một khoảng thời gian khác nhau.
* Ngân sách đào tạo theo số giờ nút giúp giới hạn chi phí.
* Bảng AutoML sẽ ngừng đào tạo nếu không đạt được mức tăng hiệu suất đáng kể.
* Đánh giá và đo lường mô hình:
* Kiểm tra các số liệu đào tạo và thận trọng với các mô hình có vẻ quá tốt để có thể trở thành sự thật.
* Các số liệu phân loại bao gồm diện tích dưới đường cong, độ chính xác, điểm F1, ma trận nhầm lẫn và tầm quan trọng của tính năng.
* Số liệu hồi quy bao gồm lỗi bình phương trung bình gốc, lỗi phần trăm tuyệt đối trung bình và tầm quan trọng của tính năng.
* Số liệu hiệu suất trên dữ liệu thử nghiệm:
* Các số liệu được tạo trên bộ thử nghiệm cung cấp cái nhìn sâu sắc về mức độ khái quát của mô hình.
* Tùy chọn đặt ngưỡng điểm cho mô hình phân loại.
* Tùy chọn triển khai:
* Các mô hình có thể được triển khai để dự đoán hàng loạt hoặc trực tuyến.
* API có sẵn cho Java, Node.js và Python.
* Giới hạn nguồn dữ liệu cho dự đoán hàng loạt:
* Đối với Bảng BigQuery, bảng nguồn dữ liệu không được lớn hơn 100 gigabyte.
* Đối với tệp CSV, kích thước của mỗi tệp được giới hạn ở 10 gigabyte, với tổng dung lượng không vượt quá 100 gigabyte.
* Lựa chọn giữa BigQuery ML, AutoML hoặc Mô hình tùy chỉnh:
* Quyết định phụ thuộc vào thời gian sẵn có và nguồn lực.
* Nên dùng thử BigQuery ML hoặc AutoML trước do rào cản thấp.
* Nếu mô hình kết quả không đủ, hãy cân nhắc đầu tư thêm nguồn lực vào việc xây dựng mô hình tùy chỉnh.