5. DATA ANALYTICS

NEXT OnAir 2020
베스핀글로벌 세션브리핑

What's New in BigQueryML,
Featuring Demandbase

Demandbase를 특징으로하는 BigQueryML의 새로운 기능

Abhishek Kashyap
Product Manager, Google Cloud

Eric White
Principal Data Scientist, Demandbase

Leo Benkel
Senior Data Engineer, Demandbase

#BigQuery #BigQueryML

© 2020 BESPIN GLOBAL, ALL RIGHTS RESERVED.

5 Data Analytics

What's New in BigQueryML, Featuring Demandbase

Demandbase를 특징으로하는 BigQueryML의 새로운 기능

BigQueryML 새로운 기능

더 많은 머신러닝 모델을 만들 수 있도록 새로운 유형의 모델들이 BigQueryML에 추가되었다. 분류(Classification) 및 회귀(Regression) 모델 같은 경우 DNN(Deep neural networks) classifier(Tensorflow), XGBoost를 사용한 Boosted trees, 그리고 AutoML Tables이 추가되었다.

기타 모델 같은 경우 단일 SQL 명령으로 수 천개 Time series 모델을 생성할 수 있는 기능과 행렬 분해를 사용한 제품 추천 시스템을 만들 수 있다. 그리고 BigQueryML에 생성된 모든 모델들은 Google AI Platform이나 자체 Stack에 내보내기가 가능하다.

BigQuery

BigQuery는 비즈니스 민첩성을 확보하기 위해 설계된 서버리스 클라우드 데이터 웨어하우스로, 높은 확장성과 비용 효율성을 갖추고있다. BigQuery를 사용하면 실시간 분석과 예측 분석으로 유용한 정보 확보, 손쉬운 데이터 액세스 및 정보 공유와 데이터 보호 및 신뢰할 수 있는 운영이 가능하다. 구글의 제일 큰 고객사는 250 PB의 데이터를 BigQuery로 이전 했으며 최대 5 PB 데이터를 쿼리하고 있다.

하지만 해당 데이터들은 데이터 분석에 용이하지만 머신러닝을 수행하려는 경우 실제로 머신러닝과 관련없는 많은 작업이 발생한다. 잘 작동하는 데이터 샘플에 대한 모델을 프로덕션 모델로 생성하려면 ETL 파이프라인을 작성해야되고 예측을 위한 ETL 파이프라인도 수행해야한다. 그러나 더 큰 골칫거리는 데이터를 내보내고 BigQuery 외부의 VM 클러스터에서 모델을 학습시키는 순간 데이터를 정의하고 있는 거버넌스 정책이 무의미해진다.
앞으로는 BigQueryML로 간편하게 머신러닝이 가능하다.

© 2020 BESPIN GLOBAL, ALL RIGHTS RESERVED.

BigQueryML

- BigQuery에서 데이터를 이동하지 않고 머신러닝 작업 실행
- 머신러닝 모델을 SQL로 작업하여 개발 속도 증가
- 기본적인 머신러닝 작업 자동화
- 내장형 인프라, 보안 및 규정 관리

AutoML Tables

AutoML Tables을 활용 하기 위해 머신러닝 관련 변수가공 지식이 없어도 되며 어떤 네트워크 아키텍처를 사용해야하는 고민없이 비정형 데이터에서 최고의 머신 러닝 모델을 구축하기 위해 만들어짐. AutoML 테이블은 자체 UI를 통해 사용할 수 있으며 BigQueryML에 새로운 모델로 추가됨. 따라서 XGBoost 모델 또는 로지스틱 회귀 모델을 BigQueryML에서 생성 할 수 있는 것처럼 이제 동일한 SQL 인터페이스를 통해 AutoML Tables 분류자를 생성하고 머신러닝 파이프라인의 데이터로 매우 쉽게 구축 가능.

BigQuery Use Cases

✔ 리테일

- 이메일 및 광고를 통한 제품 추천
- 이메일 / 메일 및 광고를 통한 구매 기반 쿠폰 성향
- 제품 및 매장의 수요 예측
- 마케팅 분석을 위한 고객 세분화

✔ 물류

- 배송 시간 추정
- 유지 보수 예측
- CPG(Consumer packaged goods)의 위치 별 재고

✔ 금융

- 소액 거래 은행 또는 보험 회사를 위한 고객 LTV(주택담보대출비율)
- 대출 신청 불이행 예측 (보험 위험)
- 신용카드 사기 감지
- 자금 세탁 검사에 대한 거래 중지 예측

© 2020 BESPIN GLOBAL, ALL RIGHTS RESERVED.

✔ **미디어, 통신사, 하이테크**

- 앱을 통한 콘텐츠 추천
- 이메일 및 광고에 대한 구독 성향
- LTV (고객평생가치) 예측
- 마케팅 캠페인에 대한 이탈 예측

Price

- 정액제 (예약) 및 Flex Slots 고객
  - Slots을 예약해 사용: Cloud AI Platform 작업은 자동으로 슬롯으로 변환된다.
- On-demand 고객
  - BigQuery를 통해 청구되는 Cloud AI Platform 비용
  - BigQuery 기본 모델 학습 비용 $ 250 / TB (미국 및 EU)

Demandbase

샌프란시스코 (HQ), 뉴욕, 시카고, 시애틀, 오스틴, 런던에 직원 300명 이상이 있으며 VC 및 전략적 투자자가 1억 달러 이상 투자한 회사. 핵심 솔루션은 계정 기반 마케팅 플랫폼, 타겟팅, 참여 및 전환 (SaaS)과 특허권계정 식별 및 AI 기술 서비스 제공.

Demandbase Use Case 1: Digital Advertising

Demandbase의 Targeting 솔루션은 AI를 활용하여 강한 구매의사를 가진 회사를 식별하는 계정 기반 광고 플랫폼이며 지출, 클릭률, 참여도, 계정 도달 범위 등을 포함한 측정 항목을 제공한다. 매월 4300억 개의 경매, 분당 5만 개의 입찰을 진행한다.

처음에는 로컬 jupyter notebook에서 모든 데이터 사이언스를 수행했음. 즉, 분석, 훈련, 검증 등을 위해 데이터를 bigquery에서 python notebook으로 내보내야 한다. 데이터 크기, CPU, 메모리 외에도 기타 제한사항들이 많아 데이터를 다운 샘플링을 해야됐고 모델의 보정과 정확도에 영향을 미쳤다.

하지만 데이터가 이미 Bigquery에 있어서 이러한 문제를 해결하기 위해 BigQueryML를 사용하기로 결정하였다. BigQueryML에서 제공하는 기능들인 로지스틱 회귀, DNN, XGBoost Trees, Model Validation을 활용해 머신러닝을 효과적으로 하고 있다.

© 2020 BESPIN GLOBAL, ALL RIGHTS RESERVED.

Demandbase Use Case 2: Account Selection

Demandbase의 Account Selection은 AI를 사용하여 마케팅 담당자가 테라 바이트의 B2B signal 및 intent 데이터를 기반으로 잠재 고객을 발굴하고 target 계정을 만들 수 있도록 한다. 또한, 고객이 계정 목록을 새로 고칠 때 "푸시 버튼"트레이닝을 지원함. 초기에는 on-demand 클러스터를 사용하여 Spark에서 개발된다.

Spark Random Forest 모델을 BigQueryML 로지스틱 회귀로 대체했으며 일부 데이터 소스를 BigQuery로 이전 하였다. 변수가공같은 경우 BigQuery의 쿼리 문으로 작성해야되는 문제가 있었지만 '쿼리 그래프로'으로 해결 함. 그래프로 구성된 BigQuery 쿼리로 데이터 파이프라인을 구성한다. (데이터 추출, 변환,로드 및 변수가공과 관련된 쿼리 포함). 각 쿼리는 이전 테이블을 사용하여 새 테이블을 만들고 결과가 캐시되므로 파이프라인 중간에서 파이프라인을 다시 재개 할 수 있다.

BigQueryML을 도입함으로써 실패율이 급격히 하락했으며 분석을 위한 시간 및 비용이 10배 이상 감소하였다.

베스핀글로벌 인사이트

그동안 BigQuery는 구글 클라우드 플랫폼(GCP)에서만 쓸 수 있었는데 앞으로는 아마존웹서비스(AWS)나 마이크로소프트 애저 같은 다른 퍼블릭 클라우드 서비스는 물론, 기업들이 자체 운영하는 데이터센터에서도 쓸 수 있게 됩니다. 그만큼 많은 사람들이 BigQuery의 성능을 인정하고 있고 BigQuery로 많은 데이터를 분석하고 있습니다.

대부분 기업들은 데이터 분석을 중요시 하고 있으며 데이터 가공, 저장 및 시각화를 통해 새로운 비즈니스 인사이트를 얻고 있습니다. 하지만 아직도 많은 기업들이 가공한 데이터를 머신러닝을 적용하고 있지는 않습니다. 머신러닝을 해야하는 대표적인 이유는 바로 데이터 자산화입니다. 비즈니스를 하면서 누적되고 있고 저장비용만 나오는 데이터는 자산이 아닌 빚입니다. 하지만 데이터를 머신러닝을 통해 예측, 추천, 자동화 등을 하면 데이터 빚을 자산으로 전환시켜 새로운 비즈니스 모델을 창출할 수 있습니다.
그리고 기업들이 머신러닝을 적용해야 더욱 경쟁력있는 회사로 발전할 수 있다고 생각합니다.

베스핀글로벌은 Google Cloud를 가장 잘 아는 전문가이며, Google Cloud의 프리미어 파트너이자,
국내 최초 Google Cloud의 MSP(Managed Service Provider)입니다.
베스핀글로벌에서는 클라우드 문의나 Google Cloud 관련 무료 컨설팅을 진행하고 있습니다.
아래 문의로 편하게 연락주시기 바랍니다.

문 의 사 항 | 베스핀글로벌 구글사업부 sales.google@bespinglobal.com 070-7931-9600

참고 웹사이트 | https://cloud.withgoogle.com/next/sf/

© 2020 BESPIN GLOBAL, ALL RIGHTS RESERVED.