데이터 과학

2011280050 컴퓨터학과 이세훈
2015130429 심리학과 정예람
2015130540 사회학과 조민서
2015131102 중어중문학과 강수진
2015170659 기계공학부 임한동

Airbnb New User Bookings

Problem

Data exploration을 통해 Airbnb 사용자 데이터셋에서 인사이트를 찾고, Airbnb 미국 신규 사용자의 첫 번째 booking 국가를 예측하는 모델을 만든다. 최종 성능평가는 Competition에서 정해진 척도인 NDCG(Normalized Discounted Cumulative Gain)@5를 통해 이루어지며, 따라서 각 사용자에 대해 1순위부터 5순위까지 다섯 국가를 예측한다.

Dataset

데이터셋명: Airbnb New User Bookings

데이터셋 출처: Kaggle

Airbnb New User Bookings는 3년 전에 열린 Kaggle competition에서 제공된 데이터셋이다. 기본적으로 Kaggle에서 제공된 데이터를 활용하되, feature engineering을 통해 새로운 feature들을 추가한다. 단, competition 규정에 의해 외부 데이터 활용은 원칙적으로 금지되나, 공휴일 정보 활용은 허용된다.

인스턴스 수:

-Training Data: 213,451

-Test Data: 62,096

클래스(country_destination): 'US'(미국), 'FR'(프랑스), 'CA'(캐나다), 'GB'(영국), 'ES'(스페인), 'IT'(이탈리아), 'PT'(포르투갈), 'NL'(네덜란드), 'DE'(독일), 'AU'(호주), 'NDF' (no destination found-예약하지 않음), 'other'(기타 국가). 총 12개.

피처(클래스 제외): id, date_account_created, timestamp_first_active, date_first_booking, gender, age, signup_method, signup_flow, language, affiliate_channel, affiliate_provider, first_affiliate_tracked, signup_app, first_device_type, first_browser. 총 15개.

1. id: User id

2. date_account_created: 계정이 생성된 날짜(20xx-xx-xx)
3. timestamp_first_active: 처음 사이트를 이용한 년-월-일-시간-분-초(총 14자리)
4. date_first_booking : 첫 예약 날짜. 사후적 정보이므로 예측에 사용할 수 없음.
5. gender: User의 성별
6. age: User의 나이
7. signup_method: basic/facebook/google/weibo
8. signup_flow: key to particular pages(Airbnb측에서 어떤 특정 사이트로 인해 에어비앤비에 signup하게 되었는지 감추기 위해 특정 사이트 URL을 숫자로 변환한 것 - 데이터 분석에 활용될 가능성이 없다)
9. language: 선호하는 언어

- Affiliate marketing - 웹 사이트 발행자(affiliate, publisher)가 그의 노력에 의해 파트너의 웹 사이트에 새로 방문자, 회원, 고객, 매출을 발생시키면, 그 웹 사이트 발행자는 소정의 보상을 받는 식의 마케팅 기법을 말한다. (배너, 상품링크 등을 통해 고객과 회사를 연결)

10. affiliate_channel: 어떤 방식의 마케팅 채널인지(Direct(email marketing, etc.)/Sem-brand(검색엔진마케팅 - 브랜드 키워드(특정 브랜드 이름 등)를 입력했을 때 검색 결과에 Airbnb 위치)/Sem-non-brand(검색엔진마케팅 - non-brand 키워드(특정 브랜드 이름이 포함되지 않는 키워드)를 입력했을 때 검색 결과에 Airbnb 위치)/SEO(검색엔진 최적화 - 콘텐츠 내 키워드의 위치나 개수 등의 관리를 통해 Airbnb를 웹페이지 상단에 위치시키기(콘텐츠 내 요소를 가지고 노출 가능성을 높이는 작업)), etc.)
11. affiliate_provider: 마케팅 채널을 어디서 제공하는지(basic, google, facebook, bing, craigslist,etc.)
12. first_affiliate_tracked: 사이트 가입 전에 처음 이용한 마케팅 방법(untracked, link, omg(Online Media Group), product)
13. signup_app : 어떤 app 운영체제를 통해서 사이트에 가입했는지(Web, iOS, Moweb, Android, etc.)
14. first_device_type : 어떤 device를 통해서 처음 접속했는지(Mac Desktop, Windows Desktop, etc.)
15. first_browser : 어떤 웹 브라우저를 통해서 처음 접속했는지(Chrome, Safari, Firefox, etc.)

Feature Engineering에 활용 가능한 부가적 데이터:

-Session 데이터: 사용자 로그 데이터 (10,567,737개 행)

1. user_id: 사용자의 ID (데이터셋과 join하여 사용하기 위함)
2. action: 세션에서 실행한 액션(show, index(view_search_results & view), search_results(view_search_results & click), personalize(wishlist와 관련), lookup, ajax_refresh_subtotal, etc.)
3. action_type: 액션 타입(data, view, click)
4. action_detail: 액션 상세 내역(wishlist_content_uploaded, view_search_results, etc)

5. device_type: 세션에 접속할 때 사용한 device 종류(Mac Desktop, Windows Desktop, etc.)
6. session_elapsed: 세션 경과 시간

-국가 데이터: 각 클래스 국가에 대한 통계값 (10개 행)

1. country_destination: 국가
2. lat_destination: 위도
3. lng_destination: 경도
4. distance_km: 미국으로부터의 거리(km)
5. destination_km2: 국토 면적(km^2)
6. language_levenshtein_distance: Levenshtein Distance를 척도로 삼았을 때 해당 국가 공용어와 영어와의 언어적 차이 수준

-클래스별 연령대, 성별 분포 (420개 행)

web browsing pattern에 대한 데이터가 주를 이룬다. (Session log, signup_app,, browser등등)

Tools

언어: python

다음의 여러 모델을 이용하며, 앙상블 기법 또한 활용한다.(gradient boosting(XGBoost,LightGBM), Catboost, logistic regression, random forest, support vector machines, KNN, ExtraTrees, AdaBoost)

최종 모델로는 explainable 모델을 채택하되, 딥러닝 모델과의 성능을 비교해본다. (tensorflow, keras, pytorch 중 활용)

Expected Result

어떤 결과를 예상하는가

- 가설1: 유저가 선호하는 언어와 첫 여행지의 언어가 가까울 가능성이 높다.
- 가설2: 모바일 휴대기기를 통해 접속하고 세션 시간이 적을수록 자국민으로서 여행에 신경을 덜 쓸 수 있는 미국으로 국내 여행을 갈 것이다.

- 가설3: 세션 시간이 길고 자주 접속할수록 이국적인 나라로 여행을 가야해서 신중하게 후보지를 고르는 사용자일 가능성이 높다. 따라서 (데이터셋이 미국유저로 이루어져있으므로) 미국과의 거리가 먼 곳으로 여행을 갈 가능성이 높다.
- 가설4: signup_method나 affiliate_provider가 facebook일 경우, 또는 접속 기기가 모바일기기이거나 first_affilate_tracked이 omg일 경우 SNS에 영향을 많이 받는 사람일 가능성이 높다. facebook의 특성과 주사용자의 특성을 생각해보았을 때 미국과 거리가 먼 유럽 국가, 특히 관광지가 유명해 사진 찍기 좋은 영국/프랑스/이탈리아/스페인을 첫 여행지로 택할 가능성이 높다.
- 가설5: 유저의 first_device_type을 이용해 연령의 missing value를 채워서 학습시키면 더 좋은 결과를 얻을 수 있을 것이다. 예를 들어, mac 데스크탑을 이용한 유저는 연령이 낮을 것으로 추측할 수 있다.
- 앙상블 기법을 사용한 모델로 NDCG@5 약 0.85-0.88의 성능을 가질 수 있을 것이다.

web browsing pattern을 통해 User의 특성을 파악해보고 이에 적합한 머신러닝 모델을 찾는다.

Role

1. 기존의 유사한 연구나 프로젝트 조사
2. Data preprocessing - identify and handle missing values(drop/replace), data normalization(data standardization), data formatting(common standard of expression), binning(transforming numerical variables into categorical bins)
3. Exploratory Data Analysis & Visualization
4. Model building & training & refinement - model selection & identifying overfitting/underfitting & calculate RSE(in-sample evaluation), Grid Search
5. Report, ppt, presentation

세훈 - 기존의 유사한 연구나 프로젝트 조사, Model building & training & refinement

예람 - Exploratory Data Analysis & Visualization, Model building & training & refinement

민서 - Data preprocessing, Model building & training & refinement

수진 - Exploratory Data Analysis & Visualization, Model building & training & refinement

한동 - Data preprocessing, Model building & training & refinement