발표_0218

1. 데이터에 알맞는 분석 도구를 선택하고 그 이유가 적절한가?
2. 데이터 분석 결과를 올바르게 해석하고 명료하게 정리하였는가?
3. 인사이트를 도출해내는 과정이 논리적인가?
4. 말하는 내용이 시각화 등을 통해 직관적으로 표현되어있는가?
5. 도출된 결론이 충분한 설득력이 있는가?
6. 발표가 매끄럽게 진행되었고 발표시간을 준수하였는가?

1. 개요

저희 조는 뉴욕 에어비엔비 데이터셋을 바탕으로 발표를 준비해온 브루클린 무브팀 입니다. 제가 발표는 대학에서 여러 번 해봤지만 이런 기술적인? 발표는 처음입니다. 최대한 저희의 생각을 잘 전달해드리는것을 목표로 잡고 발표 진행해보겠습니다.

저희 발표 순서는 다음과 같습니다.
<목차 그림 필요>

1. 개요 및 문제 정의
2. 데이터셋 소개:
3. 과정
4. Day 1
5. Day 2
6. Day 3
7. 결론

주제를 어떤 것으로 선정해볼까 팀원 분들과 이야기를 많이 나누다가, 작년 여름 뉴욕에 다녀오신 팀원분이 계셨습니다. 뉴욕에 다녀온 이야기들을 간단히 들어보다가 자연스럽게 팀 주제로 뉴욕 에어비엔비 데이터셋을 선택하게 되었습니다. 그런데 데이터를 분석하면서 '이 데이터셋을 가지고 우리가 무엇을 할 수 있을까?' 라는 질문에는 쉽게 답을 찾기 어려웠습니다.

이에 뉴욕과 관련된 다양한 논의를 나누던 중, 뉴욕의 식당 데이터셋을 살펴보게 되었습니다. 데이터를 확인한 결과, 총 1만 개의 식당 중 한식당이 단 79개에 불과하다는 점이 눈에 띄었습니다. "한식이 나름 유명한데, 이 숫자는 너무 적은 게 아닐까?" 하는 의문이 들었고, 이를 바탕으로 기존 데이터셋과 식당 데이터셋을 함께 활용해 보자는 결론을 내리게 되었습니다.

K-콘텐츠의 인기는 한국인이 체감하는 것보다 훨씬 높다고 느낍니다. 넷플릭스 순위에서 한국 작품이 상위권을 차지하는 모습을 보면 아직도 신기할 정도입니다. 이렇게 다양한 문화 콘텐츠에서 입지가 넓어짐에 따라 한식에 대한 관심도 함께 증가하고 있습니다.[1]. 전통

[1] http://www.koreatimes.com/article/20241022/1535224

한식뿐만 아니라 다양한 스타일의 한식당이 늘어나고 있으며, 이를 뒷받침하듯 미국 대상 김치 수출액도 2018년 890만 달러에서 2023년 4000만 달러로 급증했습니다.

뉴욕은 미국에서 가장 많은 관광객이 찾는 도시 중 하나이며, 관광 산업이 지역 경제에 미치는 영향이 큽니다. 따라서 에어비앤비 숙소의 인기와 한식당 입지라는 두 가지 지점에서 오늘 발표를 들어 주시면 감사하겠습니다.

2. 데이터셋 소개

이 데이터셋은 2024년 1월 5일 기준으로 작성된 미국 뉴욕 시의 Airbnb 지표입니다[2][3]. 이 자료의 경우 원저작자는 Vrinda Kallu라는 분이고, 2019년도 데이터와는 다르게 CC BY-SA 4.0 라이선스 하에 제공되고 있다는 점을 참고로 말씀 드리겠습니다. 19년도 데이터셋과 약간 다르게 침실이나 욕조 유무, 침대 개수 등 일부 칼럼이 추가되어있습니다.

다음으로 참조한 데이터셋은 2023년 작성된 Trip Advisor의 NYC 레스토랑 데이터셋 10k+ 입니다[4]. 이 데이터셋은 10397곳의 식당 데이터의 이름, 리뷰 수, 카테고리 등 6가지 칼럼을 제공합니다. (저희는 이 가운데 한식당 카테고리에 해당하는 79곳의 식당을 찾아내 활용하였습니다.)

Airbnb 데이터셋을 2024년도 문서로 활용한 이유로는 가장 최신의 데이터셋을 사용해보고 싶다는 생각도 있었지만, 같이 사용하게 된 레스토랑 데이터셋이 2023년 데이터였기 때문에 비슷한 시점의 데이터를 활용하고자 하였습니다. 게다가 19년도 이후 코로나 여파로 당시 미국을 포함한 전세계 숙박업계가 강하게 타격을 입은만큼, 데이터셋 간에 시점을 비슷하게 가져가는 것이 필요하다고 생각하였습니다.

3. 그래프 소개

<산점도 1> 이 그래프는 위도 경도를 기준으로 숙소 가격을 나타낸 그래프입니다. 일부 이상치를 제거한 이 그래프를 자세히 보시면 맨해튼과 그 주변에 비교적 가격이 높은 숙소들이 자리하고 있음이 드러납니다. 맨해튼 자체가 지대가 높기도 하지만, 주요 관광지가 맨해튼에 몰려있어서 이런 현상이 나타날 수도 있을 것 같습니다. 하지만 다음 그래프를 보시면 시사점이 하나 더 나타납니다.

<산점도 2> 이 그래프의 경우에는 월평균 리뷰 수를 기준으로 위와 같이 나타낸 그래프입니다. 리뷰 수를 기준으로 보면 위에서 나타난 중앙적 경향은 크게 나타나지 않고 거의 전 지역에 걸쳐 고루 분포되어있는 모습인 것을 알 수 있습니다. 월 평균 리뷰 수가 높다는 것은 일단 방문객수에 비례할 것이고, 그 지속력이 데이터셋의 시점까지 유지되고 있다는 뜻으로 해석 할 수 있을 것입니다. 따라서 기존 맨해튼 위주로 분포되어있는 호텔 숙박업 형태에서 지역적으로 관광객의 분포를 넓게 퍼트리는 것에 에어비엔비가 기여한다고도 말씀 드릴 수 있겠습니다. 아래 막대 그래프에서 지역별로 다시 나타내 보았습니다.

---

[2] https://www.kaggle.com/datasets/vrindakallu/new-york-dataset
[3] https://insideairbnb.com/get-the-data/
[4] Licence: CC0

<막대 1> 숙박비 차이를 막대 그래프로 나타내니 차이가 더 도드라지는 모습입니다. 맨해튼의 평균 숙박비는 다른 세 지역에 비해 1.5배? 이상 차이가 나타나고 있습니다. 간단한 조사 결과 맨해튼의 지대는 브루클린의 두 배 가량 된다고 합니다.

<막대 2> 다음 그래프를 보시면, 월 평균 리뷰수는 전과 다르게 큰 차이가 나타나지 않는 모습을 알 수 있습니다. 그나마 그 중에서 퀸스와 스태튼 아일랜드의 리뷰 수가 약간 많아 보이는 점은 눈에 띕니다. 이유가 무엇일까요? 참고로 퀸스에는 뉴욕의 공항 3곳 중 2곳이 자리하고 있다고 하고, 퀸스에는 자유의 여신상을 지나가는 페리가 무료로 운영되고 있다고 합니다. 이 외에도 다양한 해석이 가능 할 것 같습니다. 여러 요인에도 불구하고 월평균 리뷰 수가 지역별로 큰 차이가 없다는 것 또한 주목할 만한 점입니다. 저희 생각의 큰 축은 이런 겁니다. 만약 호텔 위주의 숙박업이 지속되었다면 이렇게 넓은 범위로 관광객을 움직일 수 있었을까요?

a. 최적의 위치란? 인기 숙소 주변

이런 생각들을 이어가서 저희 팀은 식당을 개업하기 좋은 위치를 인기 숙소 주변으로 생각하고 접근하게 되었습니다. 다시 말하면 인기 숙소 주변일수록 관광객의 왕래가 잦은 지역일 가능성이 높기 때문에, 인기가 높은 숙소를 기준으로 입지를 찾아보는 것이 가능할 것이라고 생각 했습니다.
뉴욕 어딘가에 가게를 개업한다고 상상해보겠습니다. 다 떠나서 위치를 정할 때 호텔의 경우라면 호텔이 그냥 일단 눈에 띄죠? 인기 호텔인지 아닌지 알아내는 것도 어렵지 않을 것입니다. 만약 호텔 주변에 가게를 연다 라고 한다면 그냥 그 호텔에서 얼마나 가까운지, 혹은 호텔에 출입하는 사람들의 눈에 얼마나 띄는지를 판단하면 될 것입니다.
반면 에어비엔비를 기준으로 가게를 개업한다는건 뭘까요? 에어비엔비 숙소는 일단 눈에 띄지 않습니다. 물론 사이트를 들어가서 다양한 판단을 해볼 수 있겠지만 기준으로 무엇 하나 정하기가 애매하다고 느껴질 것입니다.

b. 인기 숙소 선정 기준? 월 평균 리뷰 수

데이터셋에는 별점, 누적 리뷰 수, 월 평균 리뷰 수 등 인기를 구분할만한 몇 가지 척도가 있지만 저희는 그 중에서 '월 평균 리뷰 수'에 주목하여 인기 숙소를 선정하였습니다. 누적 리뷰 수 칼럼과 마지막 리뷰일자를 고민해보기도 하였지만 개업일이 제각각이라던지 적합하지 않다는 판단을 했습니다. 다른 더 많은 요인들을 변수로 활용할 수도 있겠지만, 이번 데이터셋에서 저희는 월 평균 리뷰 수를 기준으로 인기 숙소를 선정하였습니다.

c. 인기 숙소와 맛집 지역 분포 간 상관관계(맛집과 인기숙소가 상관관계가 있는지)

숙소 주변의 레스토랑과 편의시설은 숙소의 매력을 높이는 중요한 요소입니다. 기존 숙소가 없던 지역에도 에어비엔비 등을 통해 관광객들이 유입되면 해당 지역의 활성화에 긍정적인 영향을 미칩니다.대부분의 호텔이 관광지의 중심부에 위치하는 반면 에어비엔비 숙박은 지리적으로 더 분산되어 있습니다. 참고한 논문에서는 호텔만 있을 때는 불가능했을 관광객의 지출에 에어비앤비가 실제로 기여하고 있음을 연구하였습니다. 논문에 따르면 관광객이 호텔 대신 에어비엔비를 선택하는 비중이 높아짐에 따라 숙박시설의 중앙화 추세가 옅어지고 상당수의 현지 레스토랑의 고용에도 기여하고

있습니다.[5] 이와 같은 논거들을 바탕으로 저희는, 인기 숙소 주변일수록 식당의 입지를 좋게 해석하는 간단한 모델을 만들어보고자 하였습니다.

d. 한식당을 뉴욕 곳곳에 침투시키기

한 가지 변수가 있습니다. 이미 몇몇 한식당이 영업중이라는 사실입니다. 저희는 뉴욕 곳곳에 한식당을 침투시켜보자는 생각을 바탕으로, 이미 한식당이 위치하는 숙소 주변일 경우 추천도를 낮추어서 출력하도록 구상하였습니다. 모델이 저희 구상대로 추천도를 반환해준다면, 인기 숙소가 많고 주변 한식당이 적을 수록 높은 추천도를 나타낼 것이고, 인기 숙소가 없는데 한식당이 개업한 곳이라면 낮은 추천도를 나타내게 될 것입니다. 물론 모델 구성에 이미 논리적으로 한계가 느껴지고 더 다양한 값을 투입해야 참고라도 할 만한 무언가가 나오겠지만, 저희는 그래도 한 번 해보는 것에 의미를 가지고 시도해보았습니다.

<지도 그림> 을 보시면 파란 점은 월 평균 리뷰 상위 250개를 나타낸 것이고, 빨간 점은 데이터셋에 Korean 카테고리로 저장되어있는 79개 가게를 전부 나타낸 위치입니다. 저희 생각은 에어비엔비의 시장, 에어비엔비류 숙박업이라고 할까요? 그런 형태가 발달하면서 한식당의 위치도 조금 더 넓은 지역으로 진출할 수 있겠다는 것입니다. 저희가 공부하며 준비해온 방향성이 실제로 뉴욕의 여러 지역에 한식당을 진출하려 하는 사람들에게 언젠가 도움이 될 수도 있지 않을까? 상상의 나래를 약간 펼쳐가면서 진행해보았습니다.

4. 코드 설명

<코드 1: 구글 맵스 api> 식당 데이터셋에는 좌표가 없었습니다. 이 코드는 한식당의 좌표를 가져오기 위해 사용한 구글맵스 api 코드입니다. 여기 보시면 리스트 컴프리헨션으로 레스토랑 이름에 각각 뉴욕이라는 단서를 달아만 줘도 api가 뉴욕에 있는 해당 레스토랑 이름의 좌표를 돌려줍니다. 너무 신기해서 한번 소개해드렸습니다.

<코드 2: isnearby> 가지고 있는 숙소의 위도 경도와 한식당의 위도 경도를 계산하는 코드입니다. 여기서 사용한 하버사인 함수는 지구는 둥그니까 를 계산하는 함수라고 생각하시면 좋습니다. return 받는 distance는 km 단위입니다. 즉 숙소 데이터셋에 1km 이내 숙소 개수를 저장한 새 칼럼을 만든 것입니다. 저희는 주변에 한식당이 없는 쪽에 추천도를 높이고 싶었기 때문에, 카운트된 숫자에 대해 역방향으로 정규화를 진행하였습니다. 숙소 주변에 가장 많은 한식당이 있는 곳의 값은 0에 가깝고, 주변에 한식당이 하나도 없다면 1이 될 것입니다. 뒤에서 설명드릴 인기도 수치에 곱셈하여 주변에 식당이 많을 때 감점이 되는 식으로 구상하였습니다.

<코드 3: 월 평균 리뷰 수> 월 평균 리뷰 수는 많게는 70개가 넘고, 적게는 0.01인 곳도 있습니다. 이 수치에도 정규화를 진행하여 리뷰 수에 비례하게 0부터 1 사이의 값으로 저장했습니다. 이제 이 두 가지 값을 곱하여 모델에 y값으로 투입하였습니다. 다음과 같은 형태가 될 것입니다.<코드 4: target_value 계산>

<코드 4: build model> Dense 층은 케창딥 교제의 회귀모델 챕터를 참고하여 간단하게 구성했습니다. fit 하기 전에 들었던 고민거리는, +40인 위도와 -70인 경도를 그대로 x값에 투입해도 학습이 원활할까? 였습니다. 저희는 올림픽 정신으로 그냥 모델을 한 번 돌려보았습니다. 그래프는 다음과 같이 나타났습니다.

---

[5] https://www.academia.edu/101079511/The_Sharing_Economy_as_a_Local_Economic_Engine_The_Heterogeneous_Impact_of_Airbnb_on_Restaurant_Employment_Growth

<손실 그래프 1, MAE 그래프 1> 그래프만 봐도 정상적인 상황이 아니라는 사실을 금방 알았지만, 다음 그림에서

<코드 5: 첫 predict > 혹시나해서 모델에 predict를 시켜보니 서로 다른 x값, 즉 서로 다른 좌표 5개가 투입되었는데 결과가 같았습니다. 코드에 보이시겠지만 같은 지역에 대한 좌표이기 때문에 소수점 단위로 숫자가 미세하게 변동하고 있습니다. 그런데 위도의 경우 좀 전 하버사인 계산을 해보면 0.1이 변하더라도 실제 거리 상 11km 가량 움직인 결과가 되기 때문에, 그 괴리를 숫자가 담아내지 못하였다는 생각이 들어 위도 경도에도 정규화를 진행해서 다시 시도해보았습니다. 그 결과는 다음과 같습니다.

<손실 그래프 2, MAE 그래프 2> 그래프가 조금은 완만한 형태를 그리며 전과 다른 움직임을 나타내었습니다. 그래프의 움직임을 해석하려면 조금 더 공부가 필요할 것 같지만, 이에 따른 새 predict 결과는 다음과 같습니다.

<코드 6: 최종 predict> 이번에는 결과값이 다양하게 분포되어있다는 것을 알 수 있습니다. 당장은 저 숫자가 어떤 의미를 가지는지 해석하기가 어렵고 저희도 공부가 조금 부족한 부분이라고 생각합니다만 조금이라도 해석해보기 위해 월 평균 리뷰 수 최상위 3곳의 정규화된 좌표에서 소수점 네 자리에서 반올림한 좌표를 투입해 결과를 지켜보았습니다. 그 결과는 다음과 같이 나타났습니다.

<코드 7: 모델 사용 예시> 다음과 같이-저희의 해석에 따르면 인기가 가장 많은- 3곳에 대한 결과값은 자세히 보시면 리뷰 수와 역정규화한 주변 한식당 수의 곱인 타겟 밸류 칼럼과 유사하게 따라가고 있는 모습입니다.

5. 한계점

오늘 저희의 발표에서는 모델학습에 주변 한식당의 위치라는 변수를 일종의 감점 요소로서 투입하여 구성해봤습니다. 실제 뉴욕 거리에는 수많은 유인 요소 및 제한사항이 공존하고 있을 것입니다. 그래도 데이터셋에 대한 접근 방법이나 토론 과정이 유익했기 때문에 의미 있는 팀 프로젝트였다고 생각합니다.
브루클린 등지는 실제로 급속도로 성장하고 있는 지역이라고 합니다. 맨해튼에 비해 지대 기준 성장률이 매우 높다고 하고, 또 에어비엔비 등 새로운 형태의 관광객들 발길이 지속적으로 증가함에 따라 관광 1위 도시인 뉴욕에 또 다른 움직임을 만들어내고 있는 것입니다. 저희의 논리구조나 또 코드로 구현하는 과정, 결과물이 모두 좀 말도 안 되게 부족했지만, 그래도 발표를 들으시면서 뭔가 아이디어? 이런 생각을 하는 사람도 있구나? 영감?이 되신다면 만족스럽게 발표를 끝내고 오늘 좀 푹 잘것 같습니다. 데이터톤 재밌고 힘들다고 퍼실님이 말씀 하셨는데, 정말로 재밌고 힘들었습니다. 감사합니다!