1.

데이콘 주제 : 심리성향에 따른 투표여부 데이터

해당주제 : 마키아벨리즘 심리테스트를 활용하여 테스트 참가자의 국가 선거 투표 여부 예측 과제입니다

1960년대 Richard Christie라는 사람이 만들었다고 합니다.마키아벨리즘 성향이 높은 부류의 사람들은 다른 사람들과 소통하는데 있어서 보다 계산적이고 신중하게 접근하는 경향이 있다,

최근 머신러닝, 딥러닝 기술이 선거 예측에 활발히 사용되고 실제 오바마 선거캠프 당시 다양한 유권자 데이터를 활용하여 치밀한 전략을 세운 선례가 있습니다.

이번 주제를 통해 마키아벨리즘 심리테스트 데이터를 활용하여 분석함으로써, 심리적 성향을 통한 선거 예측 결과를 도출해 낼 수 있습니다.

2.

저희의 목차구성은 데이터소개, EDA, 모델링, 결론 순으로 구성되어있습니다.

3.

저희가 사용한 데이터는 다음과 같은 칼럼으로 구성되어있습니다.

QaA-QtA까지는 질문에 대한 답변 칼럼

QaE-QtE까지는 해당 질문에 답변하기까지 걸린 시간 칼럼

tp, wf, wr은 성향/단어의 의미를 아는가에 대한 답변을 수치화한 칼럼

연령, 교육 수준 등 개인특성을 나타낸 컬럼들이 있습니다.

4.

그리고 이러한 데이터 칼럼들을 유형에 따라 분류해 보았습니다.

범주형 데이터에는 모국어, 성별, 어떤 손을 주로 쓰는지, 결혼 여부, 인종, 종교, 유년기, 연령대, 교육수준의 거주 지역칼럼이

이산형 데이터에는 형제자매 수 칼럼이

순서형 데이터에는 질문에 대한 답변, 자신의 성향을 표현하는 단어, 허구 단어의 정의를 아는지, 실존하는 단어 정의를 아는지를 점수로 나타낸 컬럼

연속형 데이터에는 질문에 답변하기 까지 걸린 시간 컬럼이 해당합니다.

5.

마키아벨리즘 성향 테스트 질문이다

각 문항별로 최소 1점에서 최대 5점 선택 (1점:강한부정, 5점: 강한긍정)

질문유형에는 긍정의 질문과 부정의 질문이 있지만 일부 질문은 비식별화를 위해 Secret 처리가 되어있어 해당질문이 긍정질문의 유형인지 부정질문의 유형인지 파악이 먼저 필요했습니다.

6.

비 식별 질문의 유형를 판단하기 위해 비 식별 질문을 제외하고 상관관계분석을 실시했다.

설문문항의 경우 특정 문항은 점수가 높을수록 마키아벨리즘적인 성향을 나타내고(긍정질문), 반대 문항의 경우 점수가 낮을 수록 마키아벨리즘적 성향(부정질문)이 나타난다.

긍정 질문끼리 혹은 부정질문끼리의 상관관계는 양의 상관관계로, 긍정질문과 부정질문의 상관관계는 음의 상관관계로 나왔습니다.

7.

이 결과를 바탕으로 Secret처리 된 데이터가 긍정의 질문인지 부정의 질문인지를 다음과 같이 유추해 볼 수 있었습니다.

Secret된 데이터 중 a,d,g,i,n질문은 부정질문의 유형으로, l,p,t질문은 긍정 질문의 유형으로 유추했습니다.

8.

긍정 질문과 부정질문은 서로 상반된 의미를 갖고 있기 때문에 부정질문의 점수를 역변환하여 긍정질문의 점수와 동일한 가중치를 갖도록 조정하였습니다.

그 후 각 질문 별 점수의 평균을 내어 Score칼럼을 추가하였습니다.

9.

전처리 된 데이터를 바탕으로 EDA를 진행하였습니다.

가장 먼저 마키아벨리즘 성향여부가 투표여부와 유의미한 차이가 있다는 가설을 세워 카이제곱검정 실시하였습니다.

p-value가 0.05보다 낮으므로 마키아벨리즘 성향여부와 투표여부에 유의미한 차이가 있다는 대립가설을 채택하여 그래프에 보이는바와 같이 마키아벨리즘 성향일수록 투표참여율이 낮다는 것을 알 수 있다.

10.

다음은 투표여부와 개인 특성간의 유의미한 차이가 있다는 가설을 세우고 카이제곱검정을 실시하였습니다.

11.

카이제곱검정 결과 개인특성과 관련된 칼럼들의 p-value는 모두 0.05이하로 개인특성 데이터는 모두 투표 여부와 유의미한 차이가 있다는 대립가설을 채택할 수 있습니다.

12.

연령 데이터는 연령이 올라갈수록 투표율이 낮아지지만 예외로 10대의 경우는 투표율보다는 비투표율이 높았습닏.

교육수준 학력이 낮을 수록 투표율이 낮다

모국어 영어이면 투표율 높음

13

가족수, 성별, 사용하는 손에 따라 투표한 사람의 수, 투표하지 않은 사람의 수를 바 그래프로 시각화 하였습니다.

14.

결혼(안한 사람일 수록 투표를 안함)

인종(동양인의경우 투표를 잘하지않음)

종교:(무신론자 투표안한사람이 많다, 기독교 투표한사람이 더 많음)

유년기를 도시에서 보낸 사람들일수록 투표 안함

15.

다음은 모델링을 진행하였습니다.

모델링에 사용된 데이터는 앞서 설명드린 개인특성에 대한 칼럼과 Score칼럼만을 사용하여 진행하였습니다.

가장 성능이 좋은 모델 선정을 위해 자주 사용되는 분류 모델 중 10가지를 선정하여 계층별 K-겹 교차 검증을 통해 모델의 정확도를 계산하였습니다.

계층별 K-겹 교차 검증 (Stratified K-Fold)은 k-fold교차검증에 비해 데이터 분포의 불균형을 방지할 수 있다는 장점이 있습니다.

16.

학습곡선을 통해 각 모델을 활용하여 train시켰을때 과대적합 혹은 과소적합의 여부를 확인해 보았습니다.

선그래프 주위에 번진정도가 과대적합/과소적합

두 그래프가 모일수록 성능이 좋은 것이라 판단

17.

학습된 모델의 중요도를 시각화하였더니 아다부스트모델을 제외한 나머지 모델들은 마키아벨리즘 score보다 성별, 학벌, 연령대와 같은 개인 특성 정보가 투표 여부에 더 큰 영향을 끼친다는 것을 알 수 있었습니다.

18.

일반적으로 앙상블 모델이 단일 모델보다 정확도 측면에서 효과적이기 때문에 저희는 앞서 선정한 분류 모델 들 중 일부를 선정하여 상관관계를 분석하였고

상관관계가 높을 수록 변수의 중요도를 비슷하게 예측하는 정도.

아다부스트를 제외한 변수의 중요도를 비슷하게 예측하는 모델들을 모아서 voting앙상블 모델을 도입하였습니다.

해당 모델을 사용하여 투표여부를 예측해보았을 때 약 70%의 정확도를 보였습니다.

19.

현재 정확도 높지 않게 예측 결과가 나왔는데

이는 독립변수를 개인특성과 마키아벨리즘 Score만 적용하였고 모델 또한 아직 voting앙상블에 적용하지 못한 모델들로 인한 결과이기 때문에 추후 정확도를 높이기 위한 개선의 여지가 있다고 생각합니다.

그래서 이 프로젝트가 끝난 이후에도 정확도를 계속 개선해 나갈 예정입니다.