Titanic kaggle competition

21. 이다혜

2020. 12. 03. 목

1. 대회 개요

2. Feature Engineering

3. 사용 모델

4. 노하우 및 소감

대회 개요

타이타닉의 침몰은 역사상 가장 악명 높은 참사 중 하나입니다. 1912 년 4 월 15 일, 첫 항해 도중 타이타닉은 빙산과 충돌 한 후 침몰하였고, 이로 인해 2,224 명의 승객과 승무원 중 1,502 명이 사망했습니다. 이 비극은 국제 사회에 큰 충격을 주었고, 선박 안전 규정을 개선하는 계기가 되었습니다.

많은 사망자가 생긴 이유 중 하나는 승객과 승무원을 위한 구명정이 충분하지 않았기 때문입니다. 침몰에서 살아남는 데는 여러 요소가 있었겠지만, 여성, 어린이 및 상류층과 같은 특정 그룹의 사람들이 생존 가능성이 더 컸습니다.

대회에서 우리는 어떤 부류의 사람들이 생존 할 가능성이 높았는 지에 대해 분석을 하고, 이를 기반으로 하여 머신러닝 모델을 만든 뒤 승선한 사람들의 생존유무를 예측합니다.

Feature Engineering

사용한 데이터 : 라벨링이 된 데이터

총 10개의 Feature : Pcalss, Name, Sex, Age, SibSp, Parch, Ticket(Number), Fare, Cabin, Embarked

모델링 하기 전 Data에 대한 분석

1. Pandas를 이용해 엑셀처럼 표를 만들어서 Data 눈으로 확인하기

2. Missingno를 이용해 Null Data의 분포 확인하기

-데이터의 분포를 보고 어떻게 수정할지 고민

-아예 Feature 전체를 삭제할 것인가?

-중간치를 적용할 것인가?

Feature Engineering

3. EDA

여러 feature들을 box plot이나 dot plot등으로 시각화 도구들을 사용하여 특성을 살펴본다. 상관관계가 있을 것 같은 feature들 간의 관계를 확인한다.

Base line을 따라 필사를 하면서 Sex, Age, Pclass와 같은 Feature에 대해 분석을 진행했다.

4. Feature Engineering

전 단계에서 진행한 EDA에 따르면 Age가 모델 형성에 중요한 Feature란 것을 확인할 수 있었다. 나이가 어릴 수록 생존률이 높은 것을 확인했다. 따라서 null data를 채우는 작업을 진행했다. Train set과 test set을 합쳐서 전체 데이터셋을 활용하였다.

사용 모델

발표 수업으로 XGBoost, Light GBM, CatBoost의 사용 방법을 들었다.
초보자로서는 선생님이 제공해주신 Dietanic을 필사하며 모델 구현을 익혀보았다.

목표 : Titanic 탑승자들의 ‘생존’ 유무를 가려내는 모델을 만드는 것
-train set의 Survived 항목을 제외한 나머지 항목으로 모델을 최적화 시키고,
Test set을 통해 생존 유무를 예측한다.

노하우 및 소감

-화상 수업으로 실습이 이루어져 개인의 학습 의지가 중요

-데이터를 시각화 하는 library를 자유롭게 활용하는 능력이 중요

-눈으로 익히는 것이 아닌 손으로 필사하며 한 줄씩 이해하는 시간이 필요

감사합니다.

2020. 12. 03. 목