
머신러닝·딥러닝
문제해결 전략
공략집
(with 미니맵)

머신러닝·딥러닝 문제해결 전략

공략집

(with 미니맵)

머신러닝·딥러닝

모험의 시작

오리엔테이션

반갑습니다, 여러분!
《머신러닝·딥러닝 문제해결 전략》의 공략법을 안내드릴 금토끼입니다.

이 책은 수많은 캐글 수상자의 노트북을 수집/분석하여 여러분께 공통된 문제해결 패턴을 안내해줍니다. 총 7개의 경진대회를 이 패턴에 따라 함께 진행하면서 자연스럽게 효과적인 프로세스와 전략을 체득할 수 있게 꾸렸습니다.

머신러닝·딥러닝 문제를 하나 해결하려면 데이터 분석부터 시작하여 적합한 모델을 설계하고 최적화를 반복하는 긴 여정을 완주해야 합니다. 체계적인 프로세스를 따르더라도 몸에 익기 전까지는 도중에 길을 잃기 쉽다는 뜻입니다.

그래서 이 책은 여러분이 표류하지 않게끔 여러 장치를 마련했습니다. 공략집은 그중 하나입니다!

공략집은 다음 순서로 구성되어 있습니다.

- 오리엔테이션
- 월드맵(전체 구성)
- 미니맵(장별 구성)

오리엔테이션이 길면 지루하니 간단히 끝내고, 바로 함께 모험을 떠나봅시다.
먼저 월드맵으로 전체 구성을 살펴봐야겠네요.

월드맵

1부. 머신러닝 레벨업의 지름길, 캐글

01장
왜 캐글인가?

02장
캐글 정복 첫걸음

03장
문제해결 프로세스
및
체크리스트

04장
데이터를 한눈에 :
주요 시각화 그래프

2부. 머신러닝 문제해결

05장
다시 살펴보는
머신러닝 주요 개념

06장
경진대회 **자전거 대여 수요 예측**
머신러닝 모델링 프로세스,
기본적인 회귀 모델들

07장
경진대회 **범주형 데이터 이진분류**
탐색적 데이터 분석,
데이터 맞춤 인코딩

08장
경진대회 **안전 운전자 예측**
여러 고급 모델링 기법
(LightGBM, XGBoost, 앙상블)

09장
경진대회 **향후 판매량 예측**
다양한 피처 엔지니어링 기법

3부. 딥러닝 문제해결

10장
다시 살펴보는
딥러닝 주요 개념

11장
경진대회 **항공 사진 내 선인장 식별**
딥러닝 모델을 다루는 방법

12장
경진대회 **병든 잎사귀 식별**
유용한 성능 향상 기법

13장
데이터셋 **흉부 엑스선 기반 폐렴 진단**
훈련과 예측 단계 함수화로
활용성 높이기

월드맵

잘 살펴보셨나요? 보셨다시피 1부에서는 모험을 떠날 채비를 하고, 2부(머신러닝 던전)와 3부(딥러닝 던전) 중 선택해서 입장하시면 됩니다.

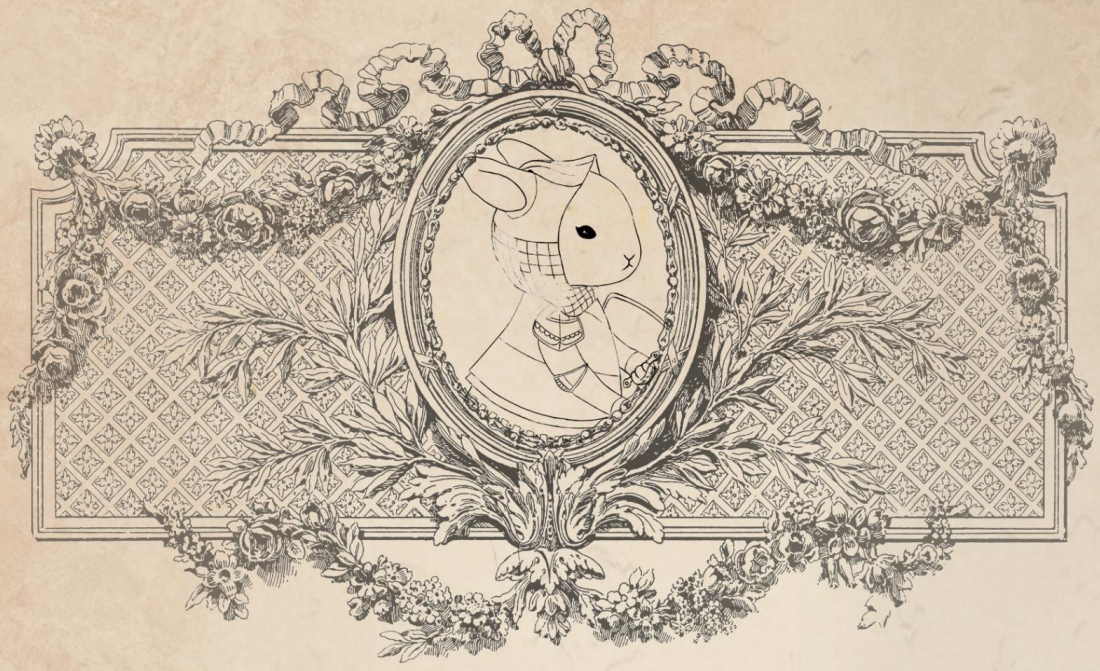
2부와 3부는 각각 주요 개념들을 정리해본 후 준비된 몇 개의 경진대회를 하나씩 공략하는 구성입니다.

여기서 중요한 사실 하나! 각 경진대회는 중점적으로 학습하는 영역이 서로 다릅니다.

예를 들어 어느 대회는 데이터 분석에 집중하고, 어느 대회는 피처 엔지니어링(특성 공학)에, 어느 대회는 고급 모델링 기법에 집중하는 식입니다.

특정 영역을 일부러 소홀히 하는 게 아니라, 해당 영역에 집중하면 좋은 성적을 낼 수 있는 대회들을 골랐습니다. 학습하는 재미와 난이도를 고려해 대회를 선정하고 배치한 것이죠!

그럼 본격적인 던전 공략에 앞서 '베이스캠프' 격의 단계인 1부로 입장해보겠습니다.



1부 머신러닝 레벨업의 지름길, 캐글

머신러닝·딥러닝 문제해결 역량을 키우는 데 캐글이 최적인 이유를 알아보고,
2부와 3부에서 본격적으로 대회를 공략하는 데 필요한 채비를 갖춥니다.

1부. 머신러닝 레벨업의 지름길, 캐글

01장
왜 캐글인가?

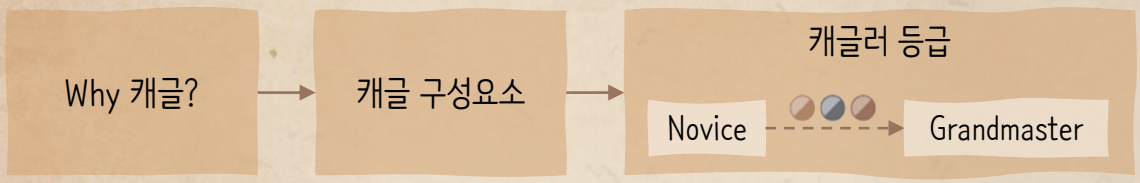
02장
캐글 정복 첫걸음

03장
문제해결 프로세스
및
체크리스트

04장
데이터를 한눈에 :
주요 시각화 그래프

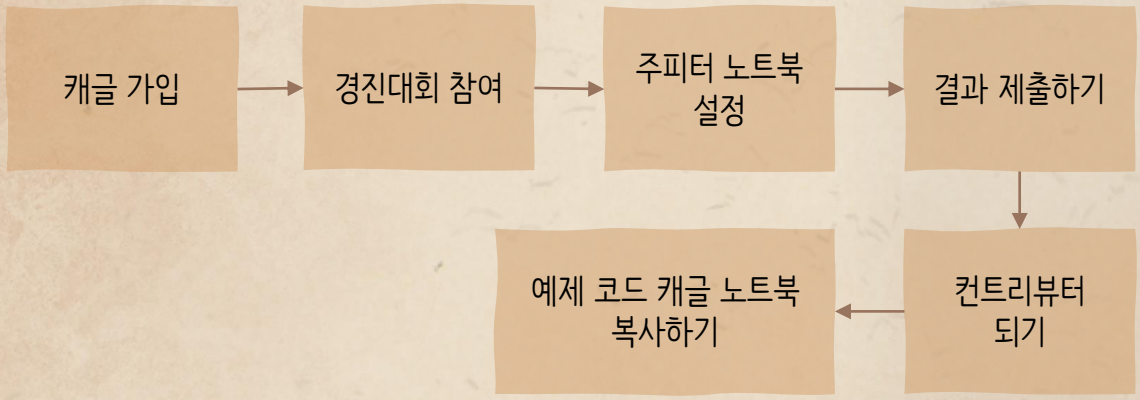
1장 왜 캐글인가?

캐글이란 무엇이고, 인공지능과 데이터 과학의 위상이 날로 높아지고 있는 오늘날 캐글이 왜 중요한지 알아봅니다. 훌륭한 머신러닝 엔지니어로 성장하는 지름길인 캐글과 친해져보세요.



2장 캐글 정복 첫걸음

캐글 가입부터 결과 제출까지 전체 프로세스를 배웁니다. 누구나 쉽게 따라 할 수 있게 캡쳐 화면으로 설명해놓았습니다.

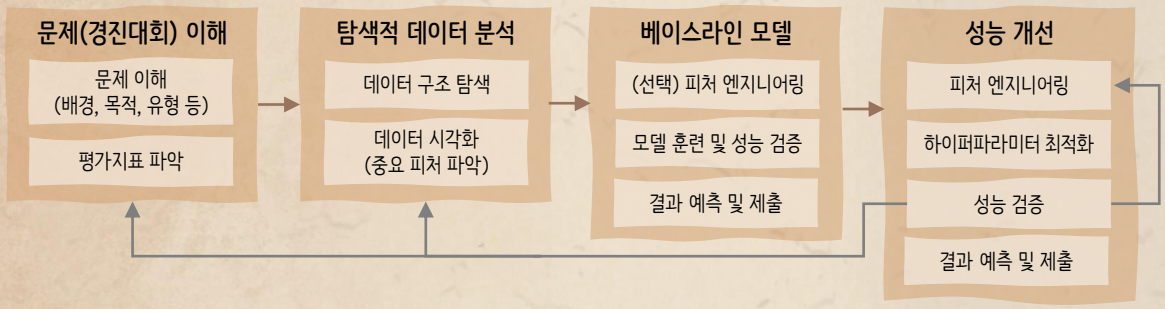


* 캐글은 웹 형태의 서비스라서 UI가 언제든 바뀔 수 있습니다. 캐글 UI가 이 책과 달라졌다면 아래 문서를 참고하세요.
https://bit.ly/3IznqWn

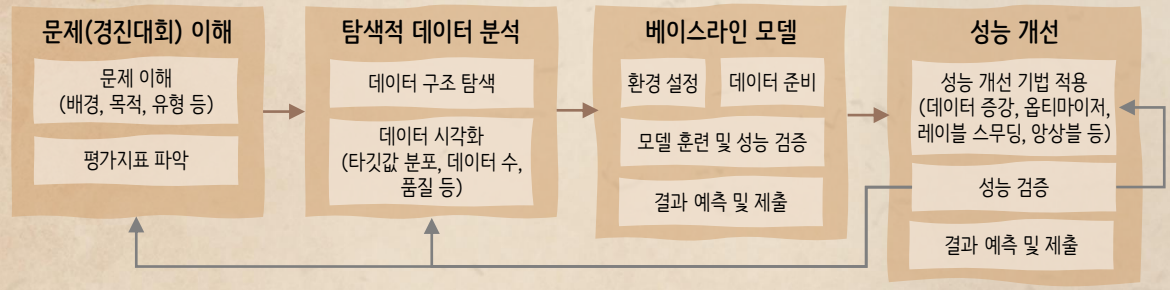
3장 문제해결 프로세스 및 체크리스트

머신러닝과 딥러닝, 두 가지 성격의 대회를 정복하는 일반적인 프로세스를 알아보고 프로세스의 각 단계에서 확인해야 할 체크리스트를 정리해봅니다.

▼ 머신러닝 경진대회 프로세스



▼ (이미지 분류) 딥러닝 경진대회 프로세스



▼ 프로세스 단계별 체크리스트

- 머신러닝 : https://bit.ly/3muJFV2
- 딥러닝 : https://bit.ly/3Bs6tLG

4장 데이터를 한눈에 : 주요 시각화 그래프

머신러닝은 데이터와의 씨름입니다. 데이터를 어떻게 이해하느냐가 모델링 전략과 예측 성능에 결정적인 영향을 줍니다. 주로 '탐색적 데이터 분석' 과정에서 수행하는 데이터 시각화는 평면적인 데이터에서 주요한 특성을 드러내는 가장 효과적인 수단입니다. 따라서 시각화 기법들을 잘 이해하고 적절히 활용하는 게 아주 중요합니다. 이번 4장에서는 다양한 시각화 기법의 개념, 효과, 구현 방법 등을 알아봅니다.

1부 정리

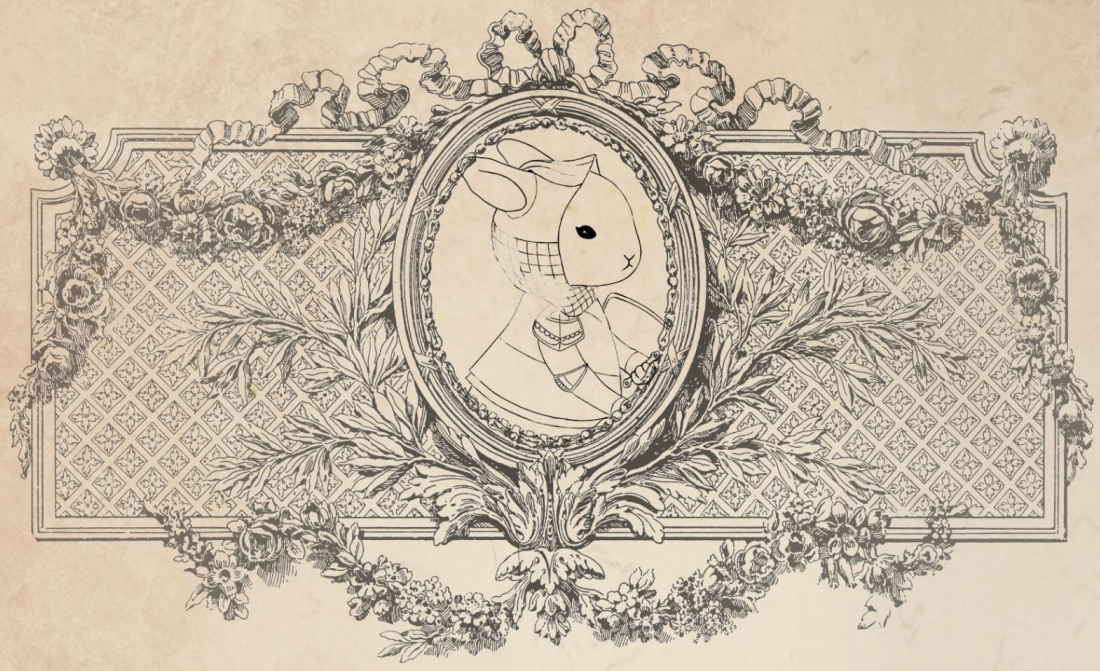
베이스캠프를 둘러봤습니다.
빠뜨린 건 없는지 점검해봅시다.

1장과 2장은 캐글 소개와 튜토리얼이니 캐글에 이미 익숙하신 분은 건너뛰어도 크게 상관 없습니다.

3장은 캐글을 경험해보신 분께도 중요합니다. 바로 이 책에서 반복 숙달할 문제해결 프로세스의 틀을 설명하기 때문입니다. 물론 이 책만의 관점은 아닙니다만, 여러 상위권 캐글러들의 공통된 패턴을 정리한 것이니 한 번씩 꼼꼼히 정리해보시면 좋을 것 같습니다.

4장은 데이터 유형을 나누고 각 유형에 유용한 시각화 기법들을 간단히 소개합니다. 2부와 3부에서(주로 2부에서) 다시 만나볼 그래프들입니다.

이상으로 모험을 떠날 채비를 다 마친 것 같군요.
그럼 머신러닝 던전부터 노크해보겠습니다.



2부 머신러닝 문제해결

머신러닝 모델을 사용하는 캐글 경진대회에 익숙해질 수 있습니다. 실제 예제를 다루면서 머신러닝 프로젝트 방법론을 터득하게 됩니다. 머신러닝 경진대회의 큰 구조는 대부분 비슷합니다. 전체적인 흐름을 파악해 머신러닝 문제에 대한 자신감을 키워보시기 바랍니다.

경진대회	문제 유형	데이터 크기	참가팀 수	난이도
자전거 대여 수요 예측	회귀	1.1MB	3,242팀	★☆☆
범주형 데이터 이진분류	이진분류	64.8MB	1,338팀	★★☆
안전 운전자 예측	이진분류	286.7MB	5,163팀	★★☆
향후 판매량 예측	회귀	96.9MB	13,613팀+	★★★

5장 다시 살펴보는 머신러닝 주요 개념

2부의 경진대회를 푸는 데 필요한 주요 머신러닝 개념들을 요약·정리해뒀습니다. 머신러닝 이론을 기초부터 차근히 설명하려는 목적이 '아니므로' 반드시 정독하실 필요는 없습니다. 궁금한 개념이 있다면 가볍게 살펴본 후 다음 장의 경진대회에 도전하시기 바랍니다. 경진대회 문제를 풀다가 언뜻 떠오르지 않는 개념이 있을 때 이번 장을 참고해주세요.

분류와 회귀

- 회귀 평가지표

분류 평가지표

- 오차 행렬 (정확도, 정밀도, 재현율, F1 점수)
- 로그 손실
- ROC AUC

데이터 인코딩

- 레이블 인코딩
- 원-핫 인코딩

피처 스케일링

- min-max 정규화
- 표준화

교차 검증

- K 폴드
- 층화 K 폴드

주요 머신러닝 모델

- 선형 회귀
- 로지스틱 회귀
- 결정 트리
- 앙상블
- 랜덤 포레스트
- XGBoost
- LightGBM

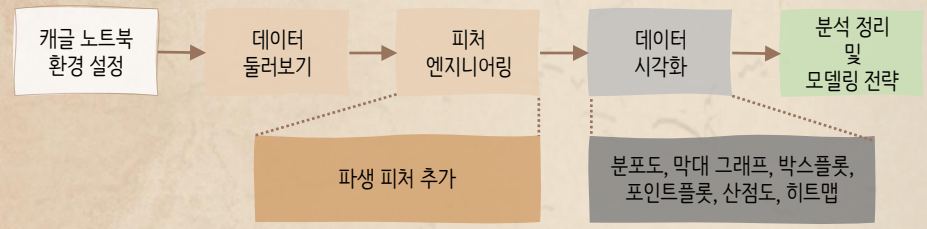
하이퍼파라미터 최적화

- 그리드서치
- 랜덤서치
- 베이지안 최적화

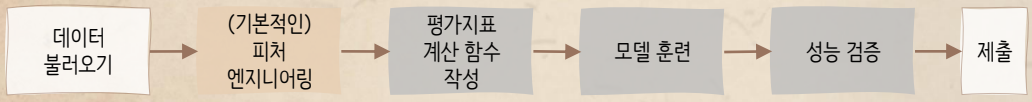
6장 경진대회 자전거 대여 수요 예측

자전거 대여 수요 예측 경진대회에 참가하여 **머신러닝 모델링 프로세스와 기본적인 회귀 모델들**을 배웁니다. 가장 먼저 경진대회 세부 메뉴를 알아보고 안내 사항을 숙지합니다. 이어서 캐글 코드를 활용해 데이터가 어떻게 구성되어 있는지 살펴보고 시각화합니다. 간단한 회귀 모델을 훈련/평가하는 방법도 알아봅니다. 마지막으로 훈련된 모델로 예측한 결과를 제출하여 순위까지 확인합니다.

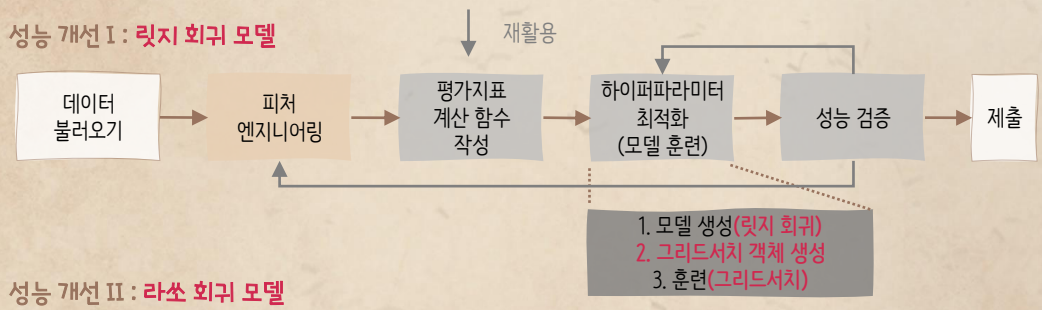
- 경진대회 이해
- 경진대회 접속 방법 및 세부 메뉴 – https://bit.ly/3ICygLg
- 탐색적 데이터 분석



- 베이스라인 모델



- 성능 개선 I : **릿지 회귀 모델**



- 성능 개선 II : **라쏘 회귀 모델**
- 성능 개선 III : **랜덤 포레스트 회귀 모델**

7장 경진대회 범주형 데이터 이진분류

이번 장에서는 범주형 데이터를 이진분류하는 대회에 참가합니다. 피처 구성을 이해하기 위해 **탐색적 데이터 분석을 자세히** 다룹니다. 모델링 단계에서는 간단한 베이스라인 모델을 만든 후 성능을 개선하여, 최종적으로 프라이빗 리더보드에서 2등을 기록하는 모델을 만들어봅니다. 이번 대회에서는 을 **데이터 특성에 따른 맞춤형 인코딩 방법**을 배울 수 있습니다.

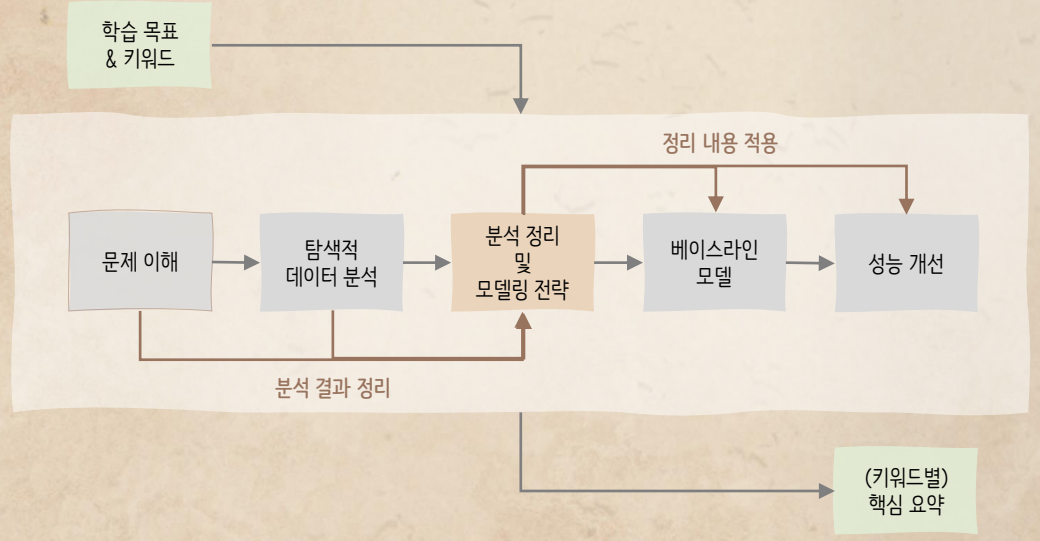
보너스! 경진대회 장들의 구성

실례하겠습니다!
중간에 끼어들어 죄송합니다! (종종 끼어들겠습니다.)

2부의 6~9장, 3부의 11~13장은 제가 미니맵을 준비했을 만큼 하나하나가 상당히 긴 마라톤과 같습니다. 분석할 데이터도 많고, 베이스라인 모델링부터 성능을 개선하기까지 뭐 하나 만만한 게 없기 때문입니다.

그래서 별도 부록 외에, 책 자체에서도 앞뒤가 유기적으로 연결되도록 다음과 같이 구성했으니 참고해주세요.

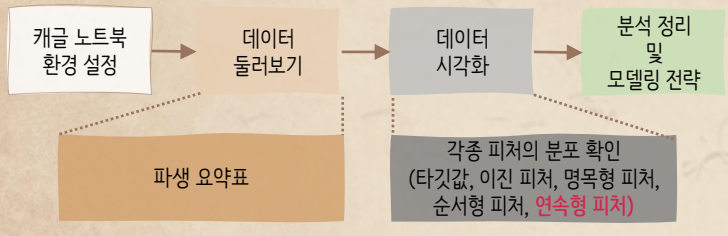
1. 주요 학습 내용을 글과 키워드로 제시 후 마지막에 다시 정리해드립니다.
2. 긴 분석 과정 후 그 결과를 정리하고, 정리한 내용이 언제 어떻게 적용되는지 이어줍니다.



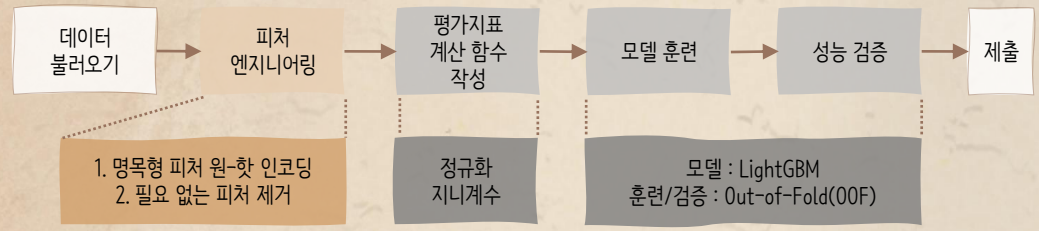
8장 경진대회 안전 운전자 예측

이번 대회에서는 실제 기업 데이트를 활용합니다. 기본 흐름을 지금까지와 같으나, 이번에는 캐글에서 실제로 많이 활용하는 **여러 가지 고급 모델링 기법**을 배울 수 있습니다. 유용한 모델들이니 잘 숙지하기시 바랍니다.

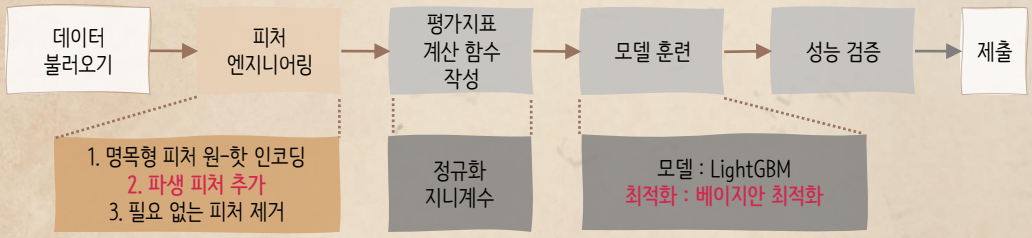
- 경진대회 이해
- 탐색적 데이터 분석



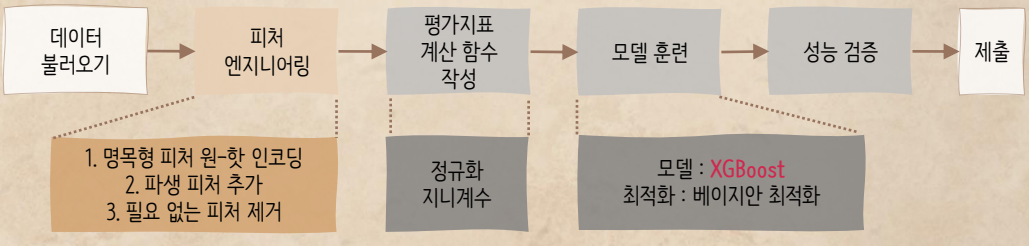
- 베이스라인 모델



- 성능 개선 I



- 성능 개선 II



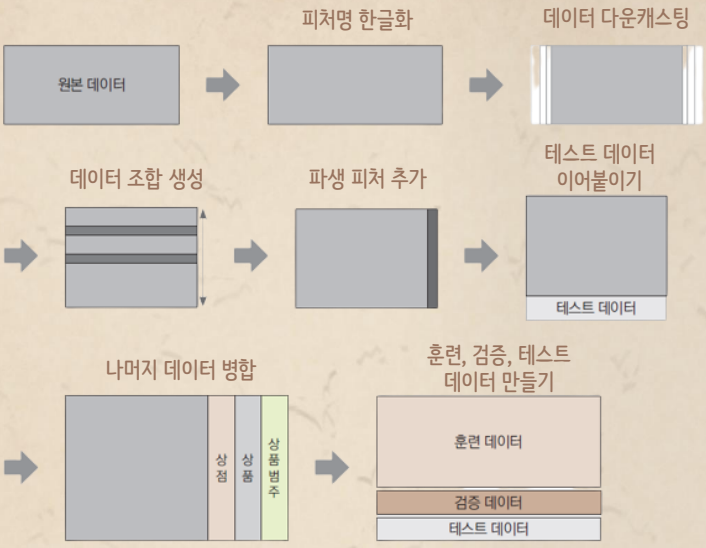
- 성능 개선 III : LightGBM + XGBoost 앙상블

9장 경진대회 향후 판매량 예측

이번 장에서는 과거 판매 데이터를 기반으로 향후 판매량을 예측합니다. 탐색적 데이터 분석은 간단하게만 다룹니다. 대신 많은 시간을 피처 엔지니어링에 할애해서 성능 향상을 위한 파생 피처를 만들어봅니다. 이 과정에서 **다양한 피처 엔지니어링 기법**을 배울 것입니다.

보너스!

9장의 중점 영역은 **다양한 피처 엔지니어링 기법**이었습니다. 실제로 베이스라인 모델링 단계에서 수행한 기본적인 피처 엔지니어링만 해도 다음과 같습니다.



하지만 이건 정말 '기본적인' 수준이고, 성능 개선 단계에 가서야 본격적인 게임이 시작됩니다. 이번 장까지 끝마치고 나면 복잡한 시계열 데이터라도 두렵지 않을 것입니다.
정말 신나는 일이군요!

2부 정리

수고하셨습니다!
이상으로 머신러닝 모델을 활용한 네 가지 경진대회를 공략해봤습니다.

6장에서는 캐글 경진대회 전반과, 베이스라인에서 시작해 모델 성능을 향상하는 일련의 절차를 배웠습니다.

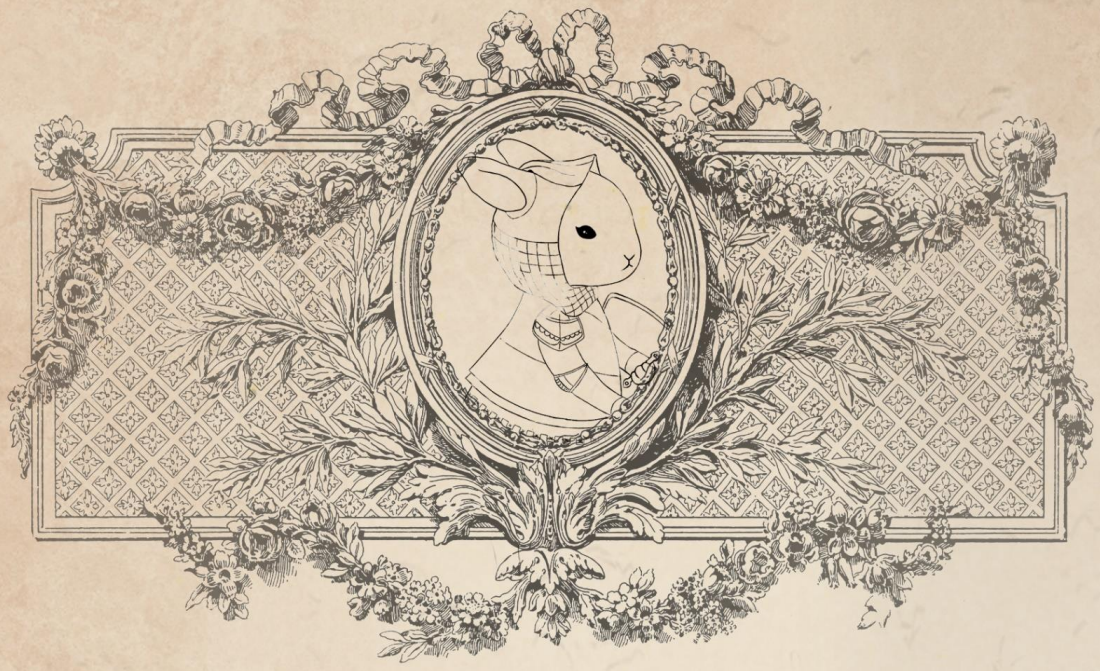
7장에서는 범주형 데이터 문제를 다뤘습니다. 탐색적 데이터 분석을 자세히 다루며 데이터 특성에 따른 인코딩 방법을 배웠습니다.

8장에서는 탐색적 데이터 분석으로 필요 없는 피처를 선별했습니다. 캐글에서 가장 많이 쓰이는 XGBoost와 LightGBM 모델 사용법도 배웠죠. 이외에도 OOF 예측, 베이지안 최적화, 앙상블 기법에 관해 배웠습니다.

마지막 9장에서는 시계열 문제에서 활용할 수 있는 다양한 피처 엔지니어링 기법을 배웠습니다.

세세한 기법들이 더 많지만 이 정도만 확실하게 알아도 머신러닝 경진대회에 참가할 자격이 충분합니다. 상위권 캐글러의 코드를 참고하면서 경진대회에 참가하다 보면 피처 중요도나 스태킹 등의 다른 기법들도 익힐 수 있습니다.

다음은 드디어 딥러닝 던전 차례입니다.



3부 딥러닝 문제해결

딥러닝 경진대회에 참가해 실력을 뽐낼 수 있도록 딥러닝 모델링을 능숙하게 하는 방법을 배웁니다. 딥러닝 프레임워크인 파이토치를 비롯해, 딥러닝 모델 구축 방법과 성능 향상을 위한 기법 등에 익숙해지시기 바랍니다.

경진대회/데이터셋	문제 유형	데이터 크기	참가팀 수	난이도
항공 사진 내 선인장 식별	이진분류	24.2MB	1,221팀	★☆☆
병든 잎사귀 식별	다중분류	785.6MB	1,317팀	★★☆
흉부 엑스선 기반 폐렴 진단	이진분류	1.15GB	1,359팀+	★★☆

10장 다시 살펴보는 딥러닝 주요 개념

3부를 진행하는 데 필요한 주요 딥러닝 개념들을 요약·정리해뒀습니다. 딥러닝 이론을 기초부터 차근히 설명하려는 목적이 '아니므로' 반드시 정독하실 필요는 없습니다. 궁금한 개념이 있다면 가볍게 살펴본 후 다음 장의 경진대회에 도전하시기 바랍니다. 경진대회 문제를 풀다가 언뜻 떠오르지 않는 개념이 있을 때 이번 장을 참고해주세요.

인공 신경망

퍼셉트론	신경망	활성화 함수 (시그모이드, ReLU, Leaky ReLU)	경사 하강법	순전파와 역전파

합성곱 신경망(CNN)

합성곱 계층	패딩과 스트라이드	풀링	전결합	전체 구조

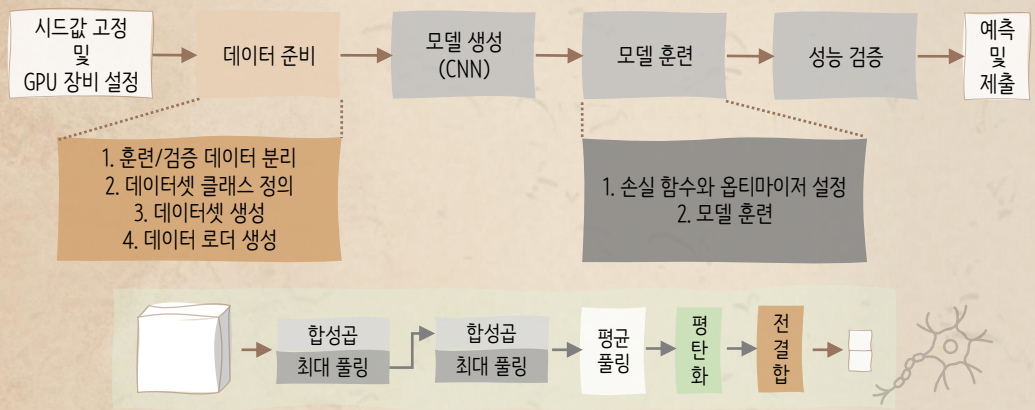
성능 향상을 위한 딥러닝 알고리즘

드롭아웃	배치 정규화 (정규화, 스케일 조정, 이동)	옵티마이저 (모멘텀, Adagrad, RMSProp, Adam)	전이 학습

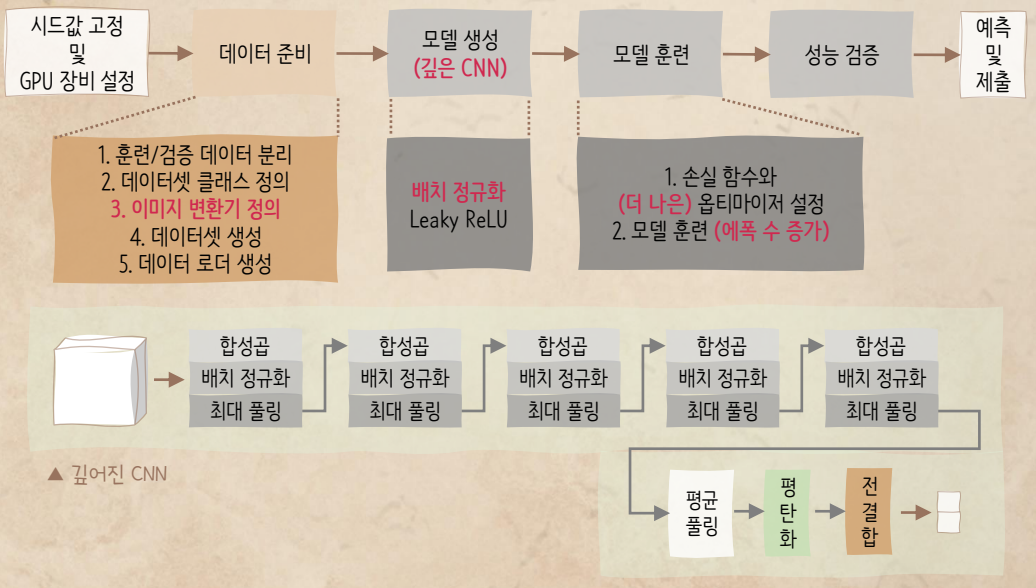
11장 경진대회 항공 사진 내 선인장 식별

이번 장에서는 딥러닝 모델을 활용한 쉬운 경진대회에 참가하여 이미지 데이터 처리 방법, 신경망 모델 설계 방법, 파이토치의 기본 활용법을 배웁니다. 첫 딥러닝 대회라서 점수 향상보다는 **딥러닝 모델을 다루는 방법** 중심으로 진행합니다.

- 경진대회 이해
- 탐색적 데이터 분석
- 베이스라인 모델



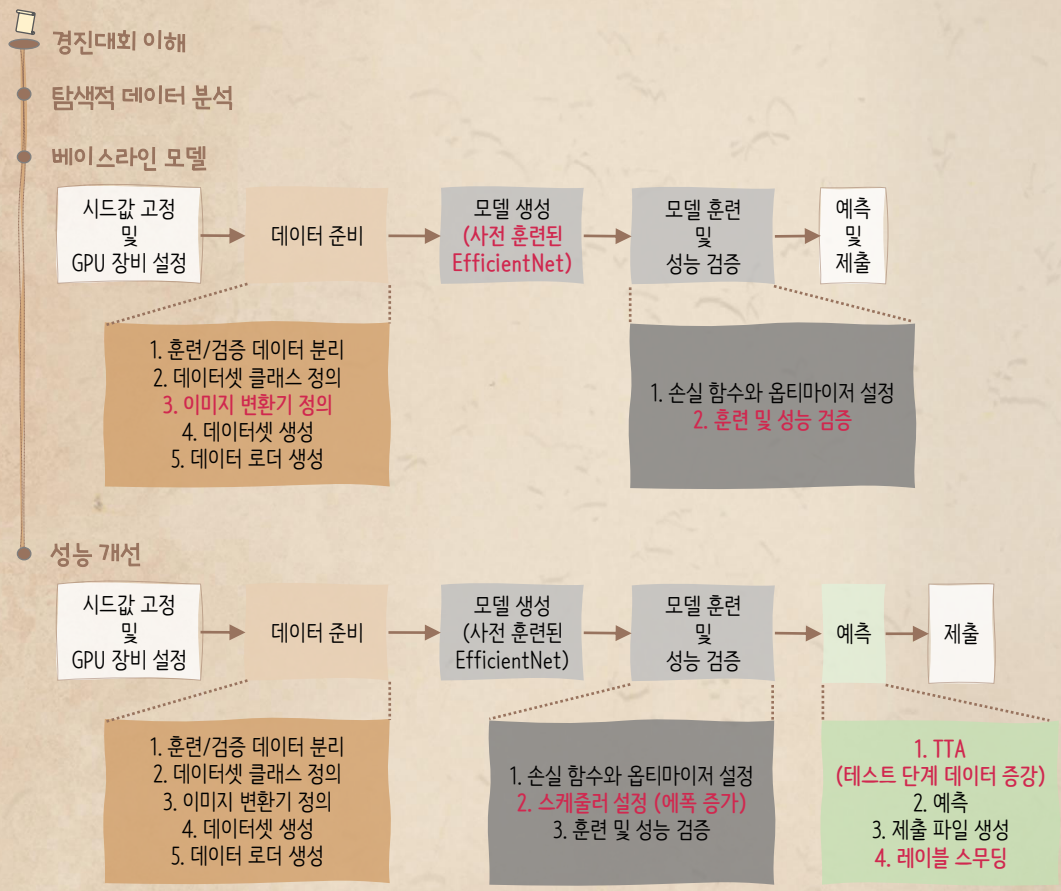
- 성능 개선



▲ 깊어진 CNN

12장 경진대회 병든 잎사귀 식별

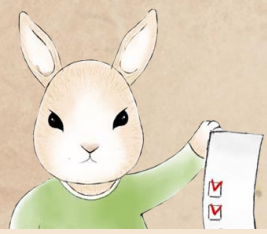
이번 장에서는 다중분류 문제를 풀어보며 몇 가지 **유용한 성능 향상 기법**을 학습합니다. 구체적으로는 사전 훈련 모델 사용법, 새로운 이미지 변환기인 Albumentations 사용법, 테스트 단계 데이트 증강 기법(TTA), 레이블 스무딩 기법 등을 다룹니다. 전체적인 모델링 절차는 앞 장과 비슷하면서 성능 향상 기법이 추가되었으니 비교하면서 학습해보세요.



13장 데이터셋 흉부 엑스선 기반 폐렴 진단

이번 장에서는 다른 캐글러가 공유한 데이터셋으로 모델링 연습을 해봅니다. 이 과정에서 **훈련과 예측 단계를 함수로 묶어 활용하는 방법**을 배우고, 앞 장에서 사용한 EfficientNet을 더 살펴보겠습니다. 전체적으로는 11~12장에서 다룬 내용과 비슷하기 때문에 복습 겸 정리하면서 '확실히 내 것으로 만든다'는 마음으로 학습해보시기 바랍니다.

보너스!



다음은 13장 베이스라인 모델링 시 작성하는
훈련 함수의 뼈대입니다.

던전에서 뼈대라니... 으스스하군요!

```
def train(model, loader_train, loader_valid, criterion, optimizer,
          scheduler=None, epochs=10, save_file='model_state_dict.pth'):
    # 총 에폭만큼 반복 ❶
    for epoch in range(epochs):
        # == [ 훈련 ] ============================================== ❷
        # 미니배치 단위로 훈련 ❸
        for images, labels in tqdm(loader_train):
            # 기울기 초기화
            # 순전파
            # 손실값 계산(훈련 데이터용)
            # 역전파
            # 가중치 갱신
            # 학습률 갱신

        # == [ 검증 ] ============================================== ❹
        # 미니배치 단위로 검증 ❺
        for images, labels in loader_valid:
            # 순전파
            # 손실값 계산(검증 데이터용)

        # == [ 최적 모델 가중치 찾기 ] ============================== ❻
        # 현 에폭에서의 검증 데이터 손실값이 지금까지 중 가장 작다면
            # 현 에폭의 모델 가중치(현재까지의 최적 모델 가중치) 저장

    return torch.load(save_file) # 최적 모델 가중치 반환 ❼
```

3부 정리

정말 수고하셨습니다!
이상으로 딥러닝 던전까지, 이 책이 준비한 모든 과정을 클리어하셨습니다.

3부에서는 비정형 데이터를, 그중에서도 이미지 데이터를 분류하는 대회들을 공략했습니다.

11장에서는 CNN 모델을 직접 설계하여 파이토치로 구현해보았습니다.

12장에서는 사전 훈련 모델, 이미지 변환기, TTA, 레이블 스무딩 등 다양한 성능 향상 기법을 선보였습니다.

마지막 13장에서는 복잡하고 반복되는 코드를 구조화하여 활용성을 높였습니다.

그럼 이만, 작별 인사 드립니다!

부디 이 공략집이 《머신러닝·딥러닝 문제해결 전략》을 학습하는 데, 나아가 더 나은 데이터 과학자/머신러닝 엔지니어로 성장하는 데 조금이나마 보탬이 되었기를 바랍니다.

머신러닝·딥러닝 문제해결 전략

캐글 수상작 리팩터링으로 배우는 문제해결 프로세스와 전략

지은이 신백균
펴낸이 최현우 · **기획** 개앞맵시(이복연)
공략집 기획 이복연 · **디자인** 조수현, 이복연

책 정보 : goldenrabbit.co.kr/xxx

마지막 업데이트 : 2022-03-31

* 본 공략집의 최신본은 다음 주소에서 받을 수 있습니다.
https://bit.ly/3tPkwbE