미술품 시장가격 예측 모델

프로젝트 3. 팀 2
안희성 추성민 박준섭 김빛나

최종발표 2024. 2. 2

TOC

I. 프로젝트 개요

기획 의도 (1)

"이 작품의 **적정 시장가격**은 어떨까?"
"**얼마까지 Bidding** 하면 좋을까?"
"이 작품은 **고평가 or 저평가** 되었나?"

취향이 자산이 되다!

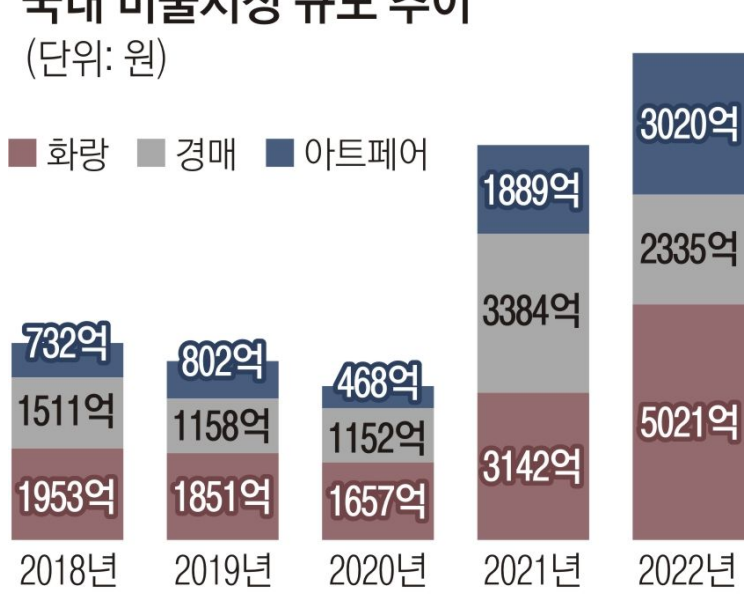
소장에 따른 심리적 만족감과
투자수익에 의한 금전적 만족감을
동시에 기대할 수 있는 미술품 구매

BUT

미술품 시장은 일반인이 이해하기 어려운 영역,

미술품 가격은 '그사세' 라는 인식

기대 효과: 미술품 가격에 대한 일반인의 예측력을 높여,
전문가에 대한 의존도와 시장 진입장벽을 낮추고자 함



〈자료: 주요 유통처 합산〉
서울신문, 2023년

기획 의도 (2)

Q) 과연 “이미지만으로” 가격 예측이 가능할까?

A) 미술품 가격이 “이미지” 그 자체만으로 결정되지 않음은 인정

BUT, 초기 가설:
대량의 이미지 학습하여 패턴 분석 ⇒ 작가/시기 추론 ⇒ 가격 예측

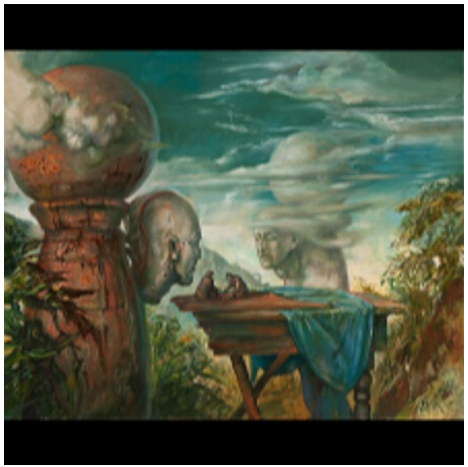
미니 테스트: 샘플 데이터(이미지 6천 건) ⇒ 성능/예측력 Not Good :(

선행연구 조사: (1) 작품 이미지 & 기타 데이터 ⇒ 가격 예측(회귀) 연구
(2) 작품 이미지 ⇒ 화가 추론(분류) 연구

가설 수정: 이미지 外 메타정보 활용 회귀모델 추가 ⇒ 성능 상승 기대

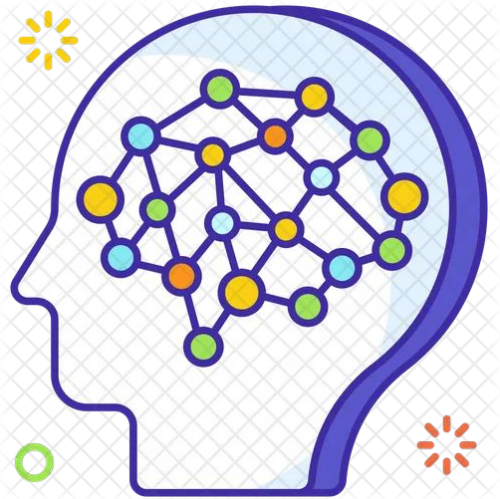
목표 산출물

Input



Samuel Bak, 2017년 작

Model



Output

How Much?

프로젝트 개요

일정		사용 자원, 기법, 모델 등
1/16 – 18	주제 선정	
1/19 – 25	데이터 수집, EDA, 전처리	BeautifulSoup, request, scikit-learn, matplotlib
1/25 – 26	샘플 모델링, 중간발표	Tensorflow, ResNet50
1/27 – 31	데이터 보강, EDA, 전처리	Tensorflow, matplotlib
1/31 – 2/1	모델 학습 및 성능 개선	Tensorflow, ResNet50, InceptionV3, vgg19
2/2	최종발표	

프로젝트 R&R

	기획	데이터 수집	EDA/전처리	모델링	PPT/발표
안희성	○	○	●	●	●
박준섭	○	●	●	○	
추성민	○	●	○	●	
김빛나	●	○	●	○	●

II. 프로젝트 수행 결과

데이터 수집

수집 데이터: 이미지, 작품명, 작가명, 제작시기, 가격(판매희망가격)

우리 모델에서 사용할 설명변수 외에 가격에 영향을 미칠만한 다른 변수들의 영향력을 줄이기 위하여 **수집 대상 데이터의 범위를 한정**함

- **고정 변수**: 회화, 고유 작품, 중간 크기(40x100cm)
- **가격(y값) 범위**: $100 – $2.5만
- **고려사항**: 균등한 y값 분포로 수집하기 위하여, 가격대별 크롤링
- **제약사항**: 웹사이트 검색 필터 1개 당 최대 3천 건 크롤링 한계

데이터 분석 및 전처리 (1)

데이터 수: 수집 **9만** → (1) 선별/전처리 후 **6.6만** → (2) 학습 **1.3만 건**

- **데이터 선별 기준**
  - (1) 가격 이상치/결측치 제거로 6.6만 건
  - (2) '**작품 데이터 개수 25개 이상' 작가만**남겨 1.3만 건

1. **가격 (labels 범위 [$100, $2.5만])**
   - 다양한 통화 → 각 단위 별 환전하여 'US$' 기준으로 통일
   - Range로 제시된 가격 → 최소값 사용
   - 이상치/결측치 제거, Standard scaling

데이터 분석 및 전처리 (2)

---

2. 작가 (약 300개 카테고리)

- *초기 가설: 작가별 작품 평균가격에 따라 작가 level 범주화?*
- 채택 방법: '**작품 데이터 개수 25개 이상' 작가만**남겨 작가별 범주화

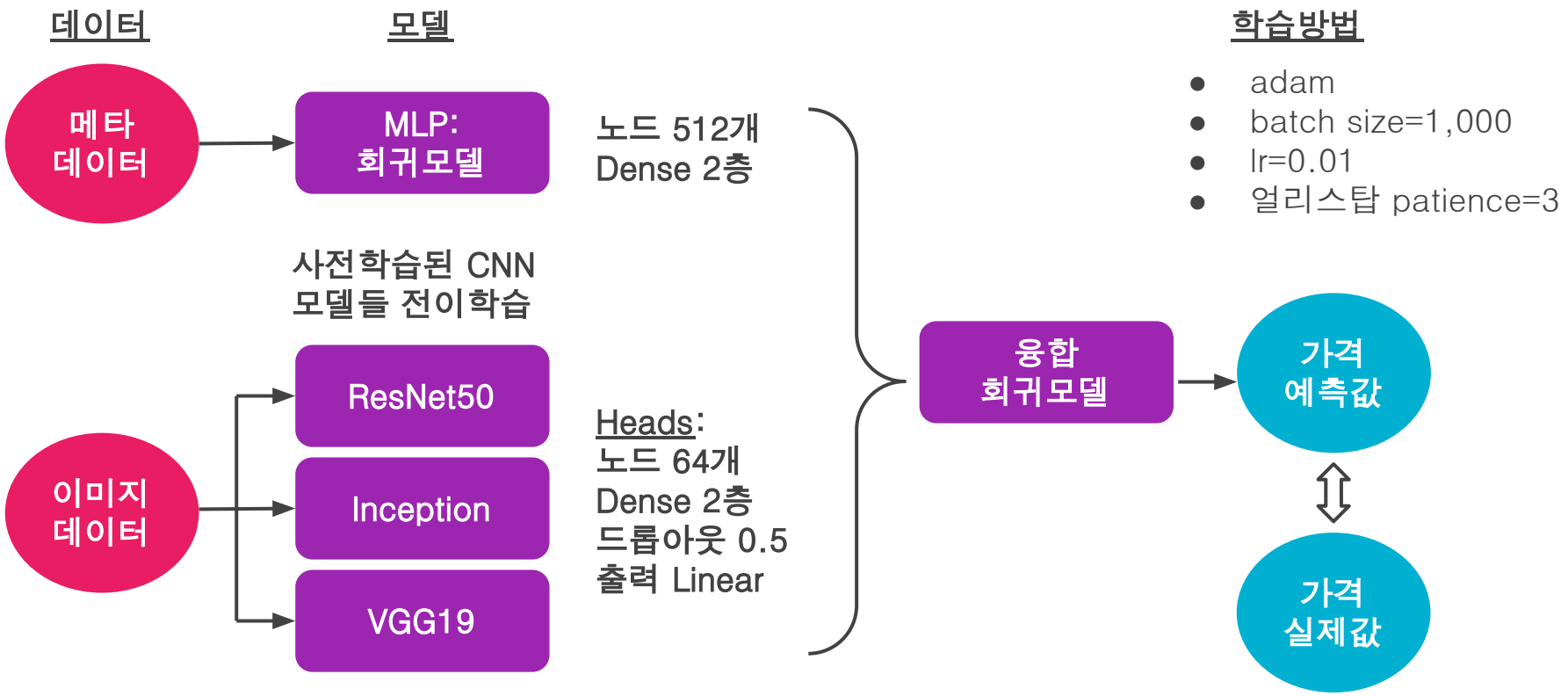
3. 제작시기 (9개 카테고리)

- 10년 단위 범주화 → 데이터 분포에 따라 '1950년대 이전'은 병합
  ⇒ 2020s, 2010s, … , 1950s, 1950 이전

4. 작품 이미지 (학습대상 데이터: 약 1.3만 건)

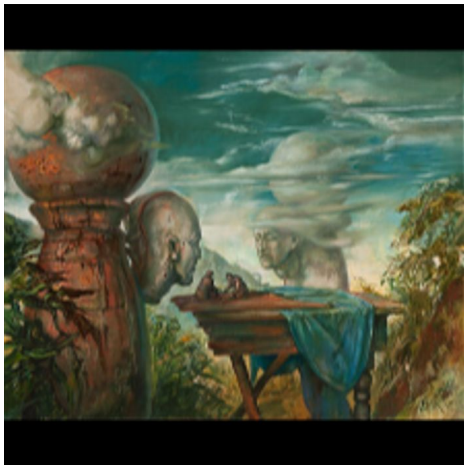
- Resizing 및 Shape 통일 (224x224x3), Scaling

모델 구조 및 학습방법


데이터
모델
메타 데이터
MLP: 회귀모델
노드 512개
Dense 2층
사전학습된 CNN 모델들 전이학습
이미지 데이터
ResNet50
Inception
VGG19
Heads:
노드 64개
Dense 2층
드롭아웃 0.5
출력 Linear
융합 회귀모델
가격 예측값
가격 실제값
학습방법
adam
batch size=1,000
lr=0.01
얼리스탑 patience=3

성능 및 실제 테스트 결과

Input



Samuel Bak, 2017년 작

Model

모델	R2
MLP	**0.518**
ResNet	3.20e-05
Inception	2.38e-06
VGG19	**0.002**
융합모델1 (MLP + CNN 3개)	0.426
★**융합모델2 (MLP + VGG)**	**0.473**

Output

예측값
$23,573
VS.
실제값
$18,000

성능 향상 시도

데이터 관점

- **메타 데이터** : EDA를 통한 outlier 제거, 데이터 수 적은 카테고리 통합
- ***이미지 데이터*** *: 고유한 창작품인 예술품 특성상, 데이터 augmentation 은 수행 안 하기로 결정*

모델 설계 관점

- **모델 구조 수정** : (1) 이미지만을 입력 데이터로 받는 가격 예측 모델 (ResNet)로 미니 테스트 → (2) 다양한 CNN 모델 성능 비교 (+Inception, VGG) & 앙상블 → (3) 최적 CNN (VGG)과 **메타 데이터 활용 MLP**를 결합한 **Multimodal 모델**

모델 튜닝 관점

- **파인 튜닝** : **CNN 모델 head 파라미터 튜닝**(드롭아웃, lr, 노드 수, batch size 등) ⇒ **BUT 성능 개선 미미**

회고 (Lessons Learned)

- **주제 선정에서 어려움을 겪었다**
  - 브레인스토밍 과정에서 나온 아이디어 중 이미지 인식보다는 객체 인식으로 해결할 문제가 많아 주제 선정 난항 ⇒ 쉽지만 뻔한 주제 ***VS* 도전적인 주제 선택**
- **데이터 수집/전처리에 시간과 리소스가 지나치게 소진되었다**
  - 다양한 모델링/성능 향상 측면에서 공부한 내용을 시도해볼 기회가 부족해 아쉬웠다
  - 주제 선정 시 **데이터 확보의 용이성**을 보다 더 고려해야할 필요성을 느꼈다
- **프로젝트 전반적으로 프로세스 관리, 커뮤니케이션의 중요성을 크게 느꼈다**
  - 여러 제약사항으로 팀원들이 동일 task를 중복하여 맡게되고 데이터 파일 개수도 늘어나며 실수와 혼선 발생 ⇒ 프로세스 차질/지연으로 연결되었다
  - **데이터 파일 버전 관리**를 철저하게 하고, **모델 학습 시마다 저장**을 반드시 해야겠다
  - 에러로 인한 코드 수정 시, 모든 해당 구간에 즉각 반영하고 주석을 달아야겠다
  - 코드에서 작명 시, 서로 이해하기 쉽고 보편적인 이름을 짓고 일관성을 유지해야겠다

개선 방향

As-Is

- **(프로세스 측면)** 모델링 단계에서 시간 부족
- **(데이터 측면)** 데이터 리소스가 단일 플랫폼에 불과하고 여러 제약이 있음: 작가 별 충분한 데이터 수 확보 어렵고, 실거래가가 아닌 판매 희망가를 시장가격 근사치(proxy)로 사용
- **(모델 측면)** 이미지 분석 모델의 단독 성능이 메타 데이터 모델 대비 현저히 낮음
- **(서비스 구현 측면)** 사용자가 이미지뿐 아니라 다른 메타 데이터도 입력해야 함

To-Be

- 충분한 데이터와 시간을 확보하여, **이미지 데이터를 sub-sampling** 한 후 앙상블 통한 성능 향상 시도
- 특히, 각 데이터 set의 가격대를 편향되도록 나눠, **서로 다른 방향으로 과적합**된 모델들을 평균 시도
- **분류 모델 시도** 및 보다 다양한 파라미터 튜닝 시도해 **최적모델 도출**
- **궁극적으로 입력 이미지 패턴 분석 만으로 가격 예측까지 직결**되는 모델 구현 및 성능 향상이 목표

참고 문헌

1. The Art of Predicting Art Auction Price (Nho & Park, 2019)
2. Painting2Auction: Art Price Prediction with a Siamese CNN and LSTM (Worth, 2020)
3. Prediction and Analysis of Artwork Price Based on Deep Neural Network (Liu, 2021)
4. Artist Identification with CNNs (Viswanathan, 2020)
5. 다변수 LSTM 순환신경망 딥러닝 모형을 이용한 미술품 가격 예측에 관한 실증연구 (Lee & Song, 2021)
6. 미술시장 활성화를 위한 AI 기반 미술품 가격 예측 및 빅데이터를 활용한 거래 트렌드 분석 기술 개발 (Kim & Lee, 2022)

감사합니다