Start Data Analysis

: 데이터 분석 프로세스를 체험할 수 있는 단계별 튜토리얼 제공

활동 배경 및 목적

데이터 분석의 중요성이 날로 증가하고 있지만, 많은 초보자들은 분석 과정에 대한 명확한 이해 없이 진입 장벽을 느끼고 있습니다.
Start Data Analysis는 이러한 초보자들을 위해 설계된 오픈소스 학습 프로젝트로, 데이터 분석의 핵심 과정을 단계별로 체험할 수 있도록 돕습니다.

이 프로젝트는 데이터 전처리, 탐색적 데이터 분석(EDA), 피처 엔지니어링, 머신러닝 모델링 등 데이터 분석의 전 과정을 따라하며 실습할 수 있도록 구성되어 있습니다. 이를 통해 초보자는 데이터를 분석하고 실제 예측 모델을 구축할 수 있는 기초 역량을 기를 수 있습니다.

주요 내용

1. **데이터 로딩 및 전처리**
   데이터를 불러오고 결측치를 처리하며, 데이터 정규화를 수행합니다.
2. **탐색적 데이터 분석 (EDA)**
   데이터를 시각화하여 분포와 상관관계를 분석하고, 데이터의 주요 특징을 파악합니다.
3. **피처 엔지니어링**
   분석에 필요한 파생 변수를 생성하고, 범주형 데이터를 인코딩하며, 불필요한 변수를 제거합니다.
4. **머신러닝 모델링**
   간단한 모델을 활용해 데이터를 학습하고 평가합니다.

활동 방법

Jupyter Notebook 기반으로, 각 단계별 학습 자료와 실습 코드를 제공합니다.

프로젝트를 클론하고, Jupyter Notebook 환경에서 학습을 진행합니다.

Python 스크립트(scripts/ 디렉토리)를 통해 각 단계의 작업을 반복적으로 수행할 수 있습니다.

기대 효과

- 데이터 분석 초보자가 분석의 전 과정을 체계적으로 경험할 수 있습니다.
- 데이터를 준비하고, 분석하며, 모델링하는 기초 역량을 갖출 수 있습니다.
- 오픈소스 프로젝트에 참여함으로써 커뮤니티와 협업의 기회를 제공합니다.

기여 방법

이 프로젝트는 누구나 기여할 수 있도록 열려 있습니다.

1. 이 레포지토리를 포크하세요.
2. 새로운 브랜치를 생성하세요.

```
git checkout -b [your-branch-name]
```

3. 변경사항을 커밋하고 푸시하세요.

```
git commit -m "Add your feature decsription"
git push origin [your-branch-name]
```

4. Pull Request를 열어주세요.

로드맵

입문자를 위한 기본 기능에서 시작해, 중급 학습자와 실무자를 위한 고급 기능과 실용적인 사례를 점진적으로 확장하는 것을 목표로 합니다. 사용자의 피드백과 커뮤니티 기여를 기반으로 지속적으로 업데이트됩니다.

1. 버전 1.0: 기초 데이터 분석

- 목표: 데이터 분석 입문자들이 전체 워크플로우를 학습하고, 간단한 머신러닝 모델을 구축할 수 있도록 기본 기능을 제공.
- 구성:
    1. 데이터 로딩 및 전처리 (결측치 처리, 정규화)
    2. 탐색적 데이터 분석(EDA) (시각화 및 상관관계 분석)
    3. 피처 엔지니어링 (범주형 변수 인코딩, 파생 변수 생성)
    4. 간단한 머신러닝 모델학습 및 평가
- 상태: 완료

2. 버전 1.1: 확장된 데이터 분석 예제 추가

- 목표: 다양한 데이터셋과 사례를 추가해 학습 내용을 확장.
- 계획:
    1. 새로운 데이터셋 제공:
    2. 고급 EDA:

        - 이상치 탐지 (Box Plot, Z-Score 활용).
        - 시간 시계열 데이터 시각화 및 분석.
    3. 데이터 전처리 자동화 스크립트 개선:
        - 중복 데이터 처리.
        - 범주형 변수 자동 처리 스크립트.

3. 버전 2.0: 고급 머신러닝 및 모델 평가 도구 추가

- 목표: 초보자를 넘어 중급 학습자가 활용할 수 있는 기능 제공.
- 계획:
    1. 머신러닝 알고리즘 추가:
        - 로지스틱 회귀(Logistic Regression)
        - Gradient Boosting (XGBoost, LightGBM)
    2. 모델 성능 평가 확장:
        - ROC-AUC Curve, Precision-Recall Curve 추가.
        - 하이퍼파라미터 튜닝(GridSearchCV) 소개.
    3. 스크립트 최적화:
        - EDA 및 머신러닝 자동화 파이프라인 제공.

4. 버전 2.1: 대규모 데이터 처리 및 시각화 개선

- 목표: 대규모 데이터셋과 복잡한 분석을 지원.
- 계획:
    1. 데이터 처리:
        - Dask, PySpark 활용한 대규모 데이터 처리 지원.
    2. 시각화 라이브러리 확장:
        - Plotly, Altair로 대화형 그래프 제공.
    3. 데이터 전처리와 EDA 통합 대시보드:

- Streamlit 기반의 대화형 데이터 분석 도구 추가.

5. 버전 3.0: 딥러닝 및 AI 모델 추가

- 목표: 머신러닝에서 딥러닝으로 확장하며, 실무에서 활용 가능한 AI 모델 제공.
- 계획:
    1. 기본 딥러닝 모델 추가:
        - TensorFlow와 PyTorch를 활용한 신경망 예제.
        - 이미지 분류(간단한 CNN), 텍스트 분류(간단한 RNN) 추가.
    2. 데이터 증강 기법 소개:
        - 이미지 및 텍스트 데이터 증강.
    3. GPU 기반 학습 환경 안내:
        - Google Colab 및 AWS SageMaker를 활용한 GPU 설정 가이드 제공.

장기 로드맵 목표

- 커뮤니티 주도형 발전.
- 학습 데이터셋과 예제 코드 기여 방안 마련.
- 실무 사례 통합:
    - 금융, 의료, 마케팅 등 다양한 산업 데이터를 활용한 실무형 프로젝트 예제 추가.