기업연계 멘토링 진행

차수명	3차시	일자/시간	2021.10.28/2pm
팀장명	이규정	기업명	신한은행
주제	데이터 이상치 처리 및 pyspark		

세부 내용

[참여인원]

4명(김윤서, 방지환, 이규정, 정희진)

[주요 진행 내용]

데이터 이상치 처리 기준 및 방법

(6개 질문)

Pyspark 소요 시간과 효율성 개선 방안

(2개 질문)

[질문/답변 내용]

1) null 값과 이상치 처리를 같은 기능으로 처리해도 괜찮을까요? 하나의 check box 로 둘 다 제거하는 방향도 괜찮은지?

- 사용자에게 선택권을 주자,
- 두 개의 버튼으로 따로 분기 처리
- 최종 유저가 소비자가 아니기 때문에 UI 고려는 유연하게 가져가도 될 것.

2) 이상치 분류 후 삭제 or 대체도 옵션 제공

- 이상치 분류 후 변경된 값을 보여주는 데 컨셉을 맞춰라!
- 원본 데이터 건들지 마라(대체값 삽입 x)
- 데이터 분석가에게 인사이트를 제공하는 목적

3) "다변량 이상치 제거" 방법을 써야되는지? 아니면 그냥 각 컬럼을 "단변량 이상치 제거" 방법으로만 하면 될지?

- 분석가에게 맡겨야 되는 부분,
- 컬럼 개별의 이상치만 제거하면 된다.
- 제거 후 다른 컬럼의 통계치 변동사항도 제공

4) 제시 방법

1. 이상치 제거 기준 후보를 옵션으로 둘지
2. 분포에 따라서 다른 기준을 **알아서** 적용해서 결과를 보여줄지
3. 혼합방법: skewed 되어있으면 IQR 적용인데 그냥 IQR or 수정 IQR 쓸지는 옵션?

- IQR or Z score 등 이상치 분류 방법을 옵션으로 둘지.
- 시스템에서 판단하여 분류해준다. 이후 그 기준을 보여줘라.
- 명확하게 기준을 설명하기 어렵다면, 빼고 그냥 추천해줘라 (우리가 임의로 선정된 알고리즘을 사용하라)
- 데이터셋 분포 특성을 고려하여 판단할 것

5) 데이터 량을 충분히 가지고 있지 않다면 결과 분석을 진행하지 못하도록 하는 장치가 필요할까요? (그렇게 적은 양의 데이터가 투입될 일이 없나요?)

- 사용자에게 맡겨라

6) 한 컬럼 내에서도 수치와 '-'가 섞여 있는 경우에는 String 으로 인식하던데, 이렇게 정제되지 않은 데이터가 넘어오는 경우도 고려해야 하나요?

- 추천을 해주고, 추천의 이유를 써주는 정도로 하면 충분하다
- 분석을 하기위해 데이터를 전처리 하기 전에 특성을 살펴보는 참고자료 이기 때문에 여기서 여러가지를 테스트 할 필요는 없다.
- '-' 가 들어있는 데이터?
- 전처리가 힘든데, 아주 간단하게는 int, float 가 섞여있는 경우가 있을 수 있다. or int 인데 excel 에서 export 해서 ,가 섞여있는 경우가 있을 수 있다. 이런 경우는 전처리 해주면 좋을 것 같다.
- '-' 까지는 사용자가 알아서 전처리 해서 특성 확인하도록 하는게 좋을 것 같다.

7) pyspark 이용해서 기본 통계치를 분석할 때 긴 소요 시간

- 분산처리 고려하지 않은 이유가, pyspark 이용했을때는 infra 가 갖춰져야 하는데 개발하거나 테스트했을때 애매할것같아서 분산처리가 안들어가면 속도측정이 의미가 없다.
- pyspark 는 그냥 작업을 진행하지 않아도 될 것 같다.
- 싱글 머신에서 pyspark 사용 시 더욱 시간이 지체된다.
- pyspark 이용은 더이상 고려하지 않아도 될 것이다.
- 특성 추출에 집중하며, 속도적인 면은 전처리에서 향상 시켜도 좋을 것 같다.

[멘토 전달사항]

정상적인 절차로 프로세스가 진행되는 것이 중요하다 사용성에 대해 스트레스 받지 말고 완주한다는 마음으로 가볍게 생각하면 좋을 것

[건의사항]