질의사항 정리

SK C&C 멘토님 질의사항

Q1-1. R&R 13페이지) 역할을 보면 웹과 모델링을 같이 한다던가, DB도 하면서 웹을 같이 한다던가. 한 사람이 각자 여러 역할을 한 걸로 보여지는데, 인원이 부족해서 그렇게 진행된 건지? 궁금합니다.

A1-1. 인원도 조금 부족했지만, 다른 조에 비해 도메인 지식을 쌓는데 시간을 길게 가졌던 이유가 가장 컸을 것 같습니다. 반도체 데이터가 저희가 쉽게 분석할 수 있는 데이터가 아니었고, 각 컬럼들이 갖는 의미도 하나하나 컸고 아무래도 쉽게 오픈되는 데이터가 아니다 보니 비공개인 정보도 굉장히 많았습니다. 각 변수들이 어떠한 의미를 가지고 있는지 추론하는 과정까지 생기면서 일정이 조금 늘어지는 바람에 결국 후반부에는 겸업으로 진행하게 되었습니다

Q1-2. 혹시 기술에 대한 호기심 때문에 더 공부를 했다던지 같은 이유가 아닌 일정의 차질을 우려해서 그렇게 진행되었을까요?

A1-2. 네 그렇습니다.

Q1-3. 저라면, 기술에 대한 호기심이 많아서라고 답변했을 것 같습니다. (웃음)

Q2-1. 또한 실시간 처리에 대해서 여쭤보고 싶은데요. 사실 실시간 대시보드 화면을 보고 제일 마음에 들었습니다. 왜냐하면 당장 제가 회사 화면에 들어가면 저런 화면이 있을 것 같아서, 그런 부분들은 정말 칭찬하고 싶습니다. 특히나 실시간으로 처리한다는 것 자체가 엄청난 트래픽을 받아내는 기술이거든요. **많은 자원을 사용하거나 대용량의 트래픽이 들어왔을 때는 어떻게 대처할 것인지**에 대해 고민을 해보신 게 있으신지 질문을 드리고 싶습니다. 앞서 활용되는 기술이라고 기술된 스택들로는 조금 무리가 있지 않을까라는 생각이 들더라구요

A2-1. 그 부분에 있어서 저도 고민을 많이 했는데, 프로젝트 내에서는 거기까지는 구현은 하지 않은 상태이구요. 트래픽이 많이 없는 상태로 구현을 할 수 있는 방향으로 노력했습니다. 제가 알고있는 파이썬 지식이나 갖고있는 최대한의 컴퓨팅 자원에 부하가 없는 적절한 정도로 구현하려고 했구요. 그러다보니 데이터 스트리밍을 프로그램으로 구현했을 때 최적화가 잘 되지 않아 서버 부하가 많이 발생하여 서버가 다운되는 등의 문제점이 발생하였고, 트러

블 슈팅의 과정을 거쳐 코딩을 최적화 하는 과정을 거치거나 서버 환경을 분화하는 작업을 거쳐 프로젝트 구현에 있어 문제가 없도록 구현을 했습니다.

Q2-2. 네, 멋있는 것 같습니다. (웃음)

Q3-1. 다음은 마지막 질문 드릴게요. 스마트 MES에 대한 질문인데요. **대부분 현장의 레거시 코드는 파이썬이 아닐 겁니다**. 그렇다면 여기와 연동할 수 있는 방안을 생각해보신 게 있으신가요?

A3-1. 좋은 질문 감사합니다. 이 부분에 대해서 질문이 들어올 것을 예상하고 있었습니다. (웃음) 원래 소스코드는 파이썬이 아니긴 한데 파이썬 라이브러리 패키지 도구 중에서 그런 시그널 정보를 받아올 수 있는 것이 있는 걸로 알고 있습니다. 그래서 설비에 부착되어 있는 센서 등에서 측정만 원활하게 이뤄진다면 라이브러리를 활용해서 사용할 수 있는 정보로 받아올 수 있는 방법이 있는 것을 확인했고, 저희가 센서가 없어서 직접 적용은 해보지 못했지만 저희가 스터디 하면서 이제 그런 시그널을 받아올 수 있는 기술들을 학습하도록 하겠습니다.

Q3-2. 가능하면, 카푸카에 대해서 공부를 해보시는게 도움이 많이 될 것 같아 보입니다.

A3-2. 좋은 조언 감사드립니다.

💡 MicroSoft 멘토님 질의사항

Q1-1. 초반에 말씀하신 것 처럼 다양한 산업에 확장이 가능할 것 같은데요, 제가 화면을 봤을 때 저희 회사에서 쓰는 시스템 같아서 되게 놀랬어요 저희 같은 경우에는 클라우드 사업도 하다보니까 데이터 센터가 전세계 100개가 넘거든요 작으면 2만대 많으면 10만대까지 있는데 거기있는 서버 하나하나가 언제 죽을지 언제 고장날지가 예측하는게 되게 중요해서 저런화면을 정말 많이 씁니다. 초 단위로 로그를 찍고하는데 역시 인제 가장 큰 문제는 고속의 대용량 데이터를 어떻게 처리할까를 궁금했었는데 답변을 잘 해주셔가지고, 그래서 한번 정도 반도체 뿐만 아니라 클라우드 산업에 붙을 수 있는 내용이긴 한데, 다른쪽으로도 활용이 더 가능하지 않을까 생각을 더 이어가면 좋을것 같구요 굳이 더 질문을 하자면 Role & Responsibillity에 질문을 하자면, 사람이 아니라 동물로 한 이유가 있나요?

A1-1. 좋은 말씀 감사드립니다. 그것은 PPT 만든 친구의 취향이였습니다….

💡 평가의원 질의사항

Q1-1. 전직 실트론에 연구소에서 일을 하면서 MES를 경험해봤습니다. 그래서인지 더 유의깊게 잘 봤습니다. 솔직하게 말씀드리면, 이렇게 예측하는 것들이 현장에서는 쉽게 일치하지가 않습니다. 설비잔여수명의 string 값은 시간이 아니라 month개념으로 나타냅니다. 시간 같은 경우에는 큰 의미가 없기 때문에… 주어진 레퍼런스가 많이 없는 상황에서 주어진 범위에서 이렇게까지라도 구현했던 것은 유의깊게 잘 봤고 교육생분들이 다룬 주제는 여전히 현업의 큰 숙제입니다. 또 하나는, **모호한 파라미터 간 관계성 29페이지** 보여주시겠어요? 이 구 형태로 나타난 문제를 해결 할 수 있는 방법은 주어진 파라미터만 저렇게 하시는 게 아니고, **PCA를 혹시 시도하셨나요**?

A1-1. 네 시도는 해보았으나

Q1-2. PCA만 쓰시면 안되고, 아마 다음에 방법에 PCA, TSME, ANN 쓰게 되면은 구분이 잘 될꺼에요 3개를 다 조합하면 뚜렷하게 구분이 갈 수 있을 것 같네요.

Q2-1. 오버 샘플링을 했나요? **데이터 언벨런스 문제** 때문에 그랬는데, 아까 설명이 이 부분이 데이터가 좀 그래서 오버 샘플링을 했다면 원래 클래스를 50 대 50으로 맞춰야 되지 않습니까? 스마트 알고리즘이나 어쨌든..

A2-2. 오버샘플링이 아니라 이상이 발생하였을 때 10분 전의 데이터를 보고 이상 근접이라는 새로운 라벨로 추가한 것입니다.

Q2-2. 그건 이제 우리 임의로 한 것이잖아요, 그렇죠? 임의대로지 그래서 보통 보면 아까 이거 이후에 설명을 했을때 비정상 데이터 너무 적어서 그랬든 이미 그전 단계에서 우리가 데이터 언벨런스 문제를 보고 Smote 알고리즘을 한 50,60 만들어놓고 학습을 시키는게 맞는데 약간 오모한 부분이 있어서 저렇게 임의대로 하시면 안돼요. 우리 10분이 아니고 15분하면 안되냐라고 질문 던지면은 답에 대한 근거를 줄 수 있나요?

A2-2. 이 10분은 저희가 임의로 지정을 한 것이고, 이러한 이유가 BtoB 산업으로 간 이유입니다. 그 산업체 데이터에서 요구하는 사항에 맞춰서 추가 샘플링 기준을 조절할 예정이였습니다.

Q2-3. 맞아요.. 그런 구간이 왜 10분이냐에 대한 질문을 예를 들면, 왜 그렇게 했느냐 그런 부분에 대해서 본인이 충분히 백그라운드를 갖고 가야하는데 어디는 15분하면 안되냐 5분하면 안되냐 이런 질문이 나올 수 있다는거지. 그런 거에 대한 근거를 스모트 알고리즘을 써가지고 50대50 맞춰보는것도 의미가 있다고 생각합니다.

A2-3. 네, 안녕하세요 이상 근접 예측 모델을 담당한 한서영입니다. 저희도 스모트 작업은 시도해 본 적 있습니다. knn imputer 작업이나, simple imputer 작업을 진행했을 때, 당연히 모델 성능 결과는 조금 더 우수했습니다. 저희 멘토님도 다행히 반도체 현직자이셨던 경험이 있으셨어서 많은 도움을 받을 수 있었는데요. 제조업에서는 이상이 발생하는 경우, 즉 이런 불균형 문제가 생기는 것이 일반적인 사례라고 하셨었고, 데이터 왜곡이 발생하는 만큼 오버샘플링 처리가 위험하다고 생각되어, 하지않는 것으로 결정했습니다.

Q2-4. 그래도 여전히 현업에서는 오버샘플링을 많이 씁니다. 말씀하신 것처럼 제조업에서의 불량 확률이 만약 50% 비율로 나타난다면 회사 다 망하겠죠. 그렇지만 어쩔수 없이 오버샘플링을 하고, 이미지 데이터 증감을 합니다. 데이터 왜곡이 불가피한 상황이고, AI를 현업에 어떻게 적용을 시키는지에 대한 방법이기 때문에 조금 더 인사이트를 얻을 수 있는 방향으로 스터디를 해 보시는 것을 조언드립니다.

💡 한화시스템 멘토님 질의사항

Q1-1. 어, 제가 궁금했던 거는 지금 비슷한 내용일 수도 있는데 10분 전을 라벨링도 하셨고 화면도 보여주셨는데 사실 설비가 고장이 나면 이걸 고치고 뭐 이런 고장을 하고 하는 과정에 프로세스 길다 보니까 사실 전조를 빨리 잡아야 하는 요구사항이 그 도메인마다 다른데 지금 반도체 공정을 하셨는데 **그 10분으로 하셨던 거, 그리고 앞에 보면 뭐 하루 전이다 수명예측을 하루 전이다. 빨간색으로 보여주셨는데 그게 도메인 공부를 하셨을 때 사유가 들어간 부분이 있는 건지 궁금합니다.**

A1-1. 10분 전의 경우에는 큰 의미를 둔 것이 아닌 이상이 일어난 범위를 좀 더 확장하고자 작업한 것이구요. progress bar의 경우에, 30일 기준으로 잡은 것은 저희의 잔여수명 데이터를 histogram으로 분포도를 그렸을 때, 75% 이상이 30일 정도의 범위가 고장이 나는 주기로 나타나는 것으로 확인이 되어, 30일이 남았을 때 Line으로 정기점검 알림을 보낸다면 30일 정도의 시간이 있으니 수리를 할 수 있을 것으로 생각했습니다.

Q2-2. 그리고 하나 더 여쭤보면 이건 그냥 제 궁금…. 해서 그런 건데 학회 데이터를 쓰셨잖아요. 그리고 논문도 두 개 정도 참고를 하셨다고 그 데이터 세트로 그 논문보다 결과가 잘 나온 건가요? 아니면 안 나온 건가요?

A2-2. 유사한 정도의 결과를 얻을 수 있었습니다. 이 학회의 논문에서도 답이 0.99의 정확도를 보였다고 하였고 저희도 0.99의 정확도를 보였다고 나왔기 때문에 유사하다고 말할 수 있을 것 같습니다

💡 쿠팡 멘토님 질의사항

Q1-1 : 그 저는 정확도 부분에서 질문사항이 있어서 질문드립니다. 제가 업체가 제조 도메인이 아니라 조금 지식이 부족할 수도 있어 드리는 질문일 수도 있는데, 저희가 보통 안돈(? 잘 들은건지 모르겠음)이란데 저희 회사도 있고, 저희들 쪽에서도 다 통용되는 개념일텐데 **이 99라는 숫자를 정말 믿어도 되는가에 대한 질문이 좀 있어서** 이 99%의 정확도면 정말 우리

는 이 데이터를 보고 설비를 점검하지 않아도 되는가 이런 궁금증이 있어서 좀 추가적으로 설명을 해주셨으면 좋겠어요.

A1-1 : 네, 저는 반도체 데이터의 특성에 의한 것이라고 개인적으로 생각이 드는데요. 반도체 공정과정이 다른 제조업의 현장과는 달리 외부 요인에 대한 영향을 최소화하고자 방진복을 입고 들어간다던지, 데이터가 조금 순수한 데이터를 얻을 수 있었어서 정확도도 잘 나왔다고 생각이 드는 부분이었습니다.