2조

학번 : 20134888, 20144701, 20144807, 20144817

이름 : 임형열, 김동진, 조재혁, 김태완

작성일자 : 2020.5.11.(월) 최초작성, **2020.5.17.(일) 최종수정**

1) 각 질문에 대한 답변 (총 17문)

Q① 도산 가능성을 통해 주식의 투자 가능성도 알아볼 수 있는지?

A : 일반적으로 도산 가능성이 낮으면 투자 위험성 또한 줄어든다고 볼 수 있습니다. 하지만 본 프로젝트의 목표는 주가를 직접 예측하는 것이 아니며, 주가의 상승과 하락 요인은 본 프로젝트에서 다루는 요소 말고도 여러 가지가 존재하므로 정확하게 알아보긴 힘들다는 점을 말씀드리고 싶습니다.

Q② 카운트 방법 사용시 기준 월 지정 방법 대비 이점(데이터 손실률)?

A : 데이터가 기준 월에 없는 기업의 경우 다른 기간에서 데이터가 존재하더라도 그 데이터를 사용할 수 없게 됩니다. 하지만 카운트 방법을 사용할 경우 기준 월에 데이터가 없더라도 전체 데이터셋에서 해당 기업의 데이터가 일정 개수 이상 존재하면 사용할 수 있다는 점에서 기준 월 지정 방법보다 손실률이 적다고 볼 수 있습니다.

Q③ 광주/전남 기업만을 대상으로 하는 것인지?

A : 원래대로라면 모든 지역의 기업에 대해 처리하여야 하지만 데이터양이 방대하여 일단은 광주/전남 기업 대상으로 진행한다는 내용을 **4주차 발표자료**에 넣었습니다. 아마도 프로젝트가 조기에 완성되지 않는 일단은 광주/전남 기업에 집중하여 분석과 예측을 진행하겠지만, 모든 지역으로 확장하여 적용시킬 수 있다는 점 또한 분명히 말씀드리고 싶습니다.

Q④ 데이터셋에는 기업에 대한 분석만 있는데, 예측률이 높다고 할 수 있는지?

A : 저희 발표자료에서 이 부분에 대한 설명이 다소 부족했던 것 같습니다. 예를 들어 총 1년도 기간의 데이터를 갖고 있다고 가정했을 때, 가장 마지막 월을 제외한 나머지 11개월의 데이터를 가지고 예측모델을 생성합니다. 그 다음 가장 마지막 월의 데이터와 직접 비교함으로써 예측률이 어느 정도인지 테스트해볼 수 있습니다.

Q⑤ 코로나19, 관련정책 같은 부정적 변수를 추가해서 예측할 수는 없는지?

A : 추가해서 예측할 수 있습니다. 하지만 코로나19같은 변수의 경우 아직 정확하게 관측된 자료가 부족하거나 아예 없다는 점에서 현 시점에서는 추가하기 힘들다고 보아야 할 것 같습니다. 관련정책의 경우 해당 데이터가 존재한다면 포함해서 예측모델을 생성할 수 있습니다.

Q⑥ 예측과 분석 결과에 대한 검증/근거는 어떻게? - 유사질문

A : 4번째 질문의 내용과 비슷해 보입니다. 전체 데이터 중에서 일부분을 테스트용 데이터로 빼놓은 다음, 나머지 데이터로 예측모델을 생성, 테스트용 데이터와 비교/대조하는 방식으로 검증을 진행할 예정입니다. 물론 이 부분에서 검증하는 것은 100% 맞다는 것이 아닌 일치하는 정도, 즉 확률에 기반한 검증이기 때문에 완전히 정확하다고 말할 수는 없겠지요.

Q⑦ 데이터의 출처가 어디인지?

중소벤처기업부(https://www.mss.go.kr/site/smba/foffice/ex/statDB/temaList.do)에서 제공하는 통계자료 중 중소기업 관련 데이터들의 모음입니다.

Q⑧ 기업의 미래 예측을 어떻게 할 것인지?

A : 4번째, 6번째 질문에 대한 답변 참고해주시면 감사하겠습니다.

Q⑨ 별도로 잡아둔 도산 가능성의 기준이 있는지, 예측 방식은 어떻게 이루어지는지?

A : 이 또한 발표자료에서 설명했던 것 같습니다. 중소기업의 평가요소 중 상위 10개 요소를 선정하여 그것을 기반으로 도산 가능성을 뽑아내려고 합니다. 예측방법은 4번째, 6번째 질문에 대한 답변 참고해주시면 감사하겠습니다.

Q⑩ 프로젝트의 최종 목표가 전국을 대상으로 하는 것인지?

A : 맞습니다. 여러 제약으로 인해 우선 광주/전남 지역을 대상으로 진행하였지만, 본 프로젝트의 최종 목표는 전국을 대상으로 하는 것입니다.

Q⑪ 전체적인 평가요소를 포함한 데이터 전처리를 하여 설계를 진행할 것인지?

A : 데이터를 가공하는 이유는 의미있는 부분만을 추려내어 사용하기 위해서입니다. 당연한 이야기지만 빅 데이터를 사용하는 최종 형태도 데이터 모두를 사용하는 것이 아니라 그것을 가공해서 유의미한 정보만 뽑아내는 것입니다. 때문에, 데이터 전처리 과정을 거치지 않고 그냥 사용하는 형태면 본 프로젝트의 의미가 없다고 봅니다. 또 전체적인 평가요소가 아닌 도산 가능성을 차지하는 비중이 가장 큰 상위 몇 개 요소를 추려내서 설계를 진행한다는 내용을 **4주차, 5-6주차 발표자료**에서 설명하였습니다.

Q⑫ 비슷한 방식과의 차이점과 일반인에게 공개된 데이터로도 충분히 가능한지?

A : 저희가 도산 가능성을 예측하는 방식은 수많은 평가요소 중 가장 영향을 많이 미치는 상위 요소 10개 정도를 추려서 그것을 기반으로 가능성을 산출하고 예측모델을 생성하는 것입니다. 또 요소 선정 과정에서 특정 은행에서 근무하시는 기업 대출 담당자의 의견을 반영했습니다. 그러나 모든 은행 기관이 똑같은 평가 기준으로 도산 위험성, 투자 가능성을 산출하지는 않습니다. 기관마다 조금씩 차이점은 있을 것이고, 이것은 신용등급 평가기관 또한 마찬가지일 겁니다. 즉 기업평가에 중점을 주는 요소가 각각 다르고, 그 때문에 위험도 수치에서 약간의 차이가 발생할 것으로 생각됩니다. 일반인에게 공개된 데이터(중소기업벤처부)로도 평가는 가능하다고 생각합니다. 물론 정확도가 매우 높은 예측까지는 불가능하겠지만요.

Q⑬ 정확히 어떤 기능이 있는 프로그램인지?

A : (데이터를 가공하여 분석/예측한 후) 기업의 도산 가능성과 그 비중이 높은 요소들을 순서대로 보여주고, 간단한 대응책까지 제시하는 기능을 구현하는 것을 목표로 하고 있습니다. 자세한 형태는 7-8주차 발표자료에서 카카오 오븐을 활용한 UI 샘플을 통해 설명해 드리겠습니다.

Q⑭ 데이터의 출처가 어디인지?

7번째 질문에 대한 답변 참고해주시면 감사하겠습니다.

Q⑮ 보여준 광주/전남 데이터의 출처가 어디인지?

7번째 질문에 대한 답변 참고해주시면 감사하겠습니다.

(광주/전남 데이터는 전체 데이터에서 가공을 통해 얻은 일부분입니다.)

Q⑯ 각자 역할이 무엇인지?
A : 팀의 개인별 담당 역할은 다음과 같습니다.
(발표자료에 이 부분을 작성하지 않았네요. 다음 발표부터는 포함하도록 하겠습니다.)

임형열 : 평가요소 산출, 파이썬을 활용한 데이터 전처리와 위험도 예측
김동진 : 파이썬을 활용한 데이터 전처리와 위험도 분석, 코드 구현 및 정리
조재혁 : 자료 수집, 코드 구현 및 정리, 웹 프레임워크 설계 및 웹 서비스 구현
김태완 : 웹 서비스 기본 형태와 웹 프레임워크 설계

각자 잘하는 부분 위주로 역할을 나누었지만, 저를 포함한 팀원 모두 정해진 역할대로만 움직이지는 않습니다. 예를 들어, 특정 부분에서 문제에 부딪히면 모든 팀원이 모여서 해결을 위한 방안을 제시하고 구현하는 등 유기적으로 움직이고 있습니다. (협업한다는 표현이 맞을까 모르겠네요.)

Q⑰ 2조 : 데이터 셋에 대한 부분을 봤는데 정확한 수치가아닌 1 매우나쁨, 5 매우좋음 이런 데이터인 것 같은데 예측에 사용할만한 데이터 인지 궁금하다 평소에는 본인은 어떤 값, 수치를 가지고 예측을 했는데 1~5번까지 크게 크게 수집된 데이터로 분석이 될지 궁금하다. ＊ 1조에서 추가로 질문한 사항

A : 일단 무슨 이야기를 하는지 잘 이해가 안 됩니다. (그럼에도 불구하고 우리가 이해한 것이 맞다면) 먼저 한 가지 예시를 들어보겠습니다. 복싱선수의 팔길이(리치) 를 길이별로 1-5구간으로 나눈, 1,2,3,4,5로 정한 데이터가 있습니다. 큼지막하게 수집된 이 데이터가 과연 승률 분석과 예측을 할 때 사용할 수 없는 자료일까요?

데이터셋을 제대로 보셨는지 모르겠지만 수치화 할 수 없는 요소들이 많구요.(예를 들어 자금사정전망이라는 요소명에서도 볼 수 있듯이 말 그대로 '전망'이지 '확률'를 의미하는 것이 아닙니다.) 또 본 데이터는 중소벤처기업부에서 제공하는 단순 통계자료이지 원래부터 특정 결과를 내기 위해 타겟으로 한 데이터가 아닙니다. 이러한 자료를 사용해도 문제가 없는 것인지는 **5-6주차 발표영상**에서도 말씀드렸지만 은행에 20년 이상 근무하셨고, 10년 넘게 중소기업의 대출심사를 담당하셨던 분(개인정보 활용에 대해 따로 동의를 구하지 않아 발표자료엔 따로 기입을 하지 않았습니다만 팀 구성원의 친인척입니다.)에게 자문을 받았으므로 어느 정도 검증되었다고 생각하구요. 처음부터 퍼센트로 표시가 되어있었다면 저희가 굳이 본 프로젝트에서 **요소 선정 및 가중치를 부여 과정(자문)**을 진행할 필요가 없었겠지요?

마지막으로 정리해서 말씀드리자면, 1 2 3 4 5 매우나쁨, 좋음 이런 식으로 1부터 5까지 '크게 크게' 조사된 데이터의 경우 상식적으로 생각해본다면 0에서 20, 20에서 40, 40에서 60 등 구간을 1부터 5까지로 나누어 기입하였다고 볼 수 있습니다. 즉 '**정확한 수치**'가 아니더라도 신뢰도는 조금 낮아지더라도 분석에는 전혀 문제가 없다는 점 말씀드립니다.

2) 8-9주차 강의에서 조언한 부분

1. PPT 목적 부분에서 '기업의 도산 가능성 분석 및 예측'이라는 내용이 '중소기업'을 타겟으로 한 주제 부분과 맞지 않으며, 왜 이런 개발환경을 사용하는지에 대한 구체적 설명이 작성되어 있지 않다.

개발환경 부분에서는 분석 및 예측결과를 웹 페이지로 출력, 통계데이터 가공, 예측모델 생성/검증이라고 쓰는 이유를 명시했습니다만, 최대한 간략화하다 보니 저희의 의견을 충분히 설명하지 못했던 것 같습니다. 이 점 반영하여 작성하도록 하겠습니다.

2. 보고서의 내용이 PPT에도 들어가야 하고 데모 영상이 필요하다

PPT에는 핵심 키워드만 넣고 설명에서 왜 키워드를 넣었는지 설명하려는 방향으로 진행하려고 했는데 아무래도 잘 전달이 안 된 것 같습니다. 보고서 내용을 조금 더 PPT에 반영하여 작성하겠습니다. 또한 데모영상은 파이썬으로 데이터를 가공하는 과정을 진행하여 따로 영상을 준비하지 않았습니다. 지난 주 부터 웹 프레임워크를 사용하여 웹 사이트를 구현해보고 있는데 금주부터는 이것을 데모 영상으로 포함하여 작성하도록 하겠습니다.

3. 설계도가 너무 광범위('Generic')하다

처음 프로젝트 설계시 설계한 내용이었고 그 뒤로 계속 세분화시켜서 작성하고 있는데 아직 보고서와 발표 자료에는 넣지 못하였습니다. 금주 발표자료부터 세분화 한 설계도를 반영하여 작성하도록 하겠습니다.

---

3) 본인 팀에서 현재 가장 어려운 부분 또는 에러가 발생했을 때 어떻게 해결하고 있는지? 본인의 역할을 구체적으로 작성해 보셔요.

위의 16번째 질문에 대한 답변처럼, 팀원 모두 정해진 역할대로만 움직이지는 않고, 특히 특정 인원을 중심으로 한 **수직적 의사결정 방식**은 더더욱 지양하고 있습니다. 특정 부분에서 문제가 나타나서 해결하지 못할 경우, 모든 팀원이 모여 해결을 위한 방안을 제시하고 가장 합리적인 방식 -모든 팀원이 납득할 수 있는- 으로 진행하는 등 토론을 통한 문제해결을 중요하게 생각하며 실제로 이러한 방식으로 프로젝트를 진행하고 있습니다.

임형열 : 제 역할은 평가요소 산출, 파이썬을 활용한 데이터 전처리와 위험도 예측 부분입니다.
평가요소 산출은 도산 가능성에 직접적으로 영향을 미칠 수 있는, 비중이 큰 상위 10개 요소 선별하고 가중치를 부여하는 과정이며 파이썬을 활용한 데이터 전처리/위험도 예측은 파이썬 3.7 기반의 아나콘다3-주피터 노트북을 활용하여 통계데이터를 가공(결측치 해결)하는 부분과 기업별 특정 기간에 해당하는 데이터 정렬, 이후 해당 데이터를 활용하여 예측모델 생성한 후, 테스트셋과 비교(검증)하는 부분입니다.

김동진 : 저는 파이썬을 이용하여 데이터 전처리를 하고 전처리된 데이터를 이용하여 프로젝트의 한 가지 항목뿐 아닌 다른 항목에도 이용할 수 있도록 가공을 하고 있습니다. 또 웹사이트 프로토타입을 김태완 팀원과 협력하여 카카오 오븐을 사용하여 설계했습니다.
조재혁 : django 웹 프레임워크를 이용하여 Model, Template, View를 작성하여 기능및 예측된 시각화된 과정을 웹에 사용자로 하여금 직관적으로 자료를 나타내는데 중점을 두고 있습니다. Model은 데이터를 수집하는 과정에 사용하고 있고, template는 팀원과 같이 웹 디자인 및 urls를 병행하여 코드를 짜고 있습니다. View는 각 action이나 from을 구성할 때 어떤 형식으로 Model에 전달할지를 정해서 코드를 짜고 있습니다.

김태완 : Django 웹 프레임워크 작업을 같이 수행하면서 bootstrap4를 이용하여 form 및 layout 디자인을 꾸미고 있으며 각 웹에 버튼에 대한 url을 관리하고 있습니다. 또 javaScript를 통하여 반응형 웹을 만드는 것을 목표로 프로젝트를 진행하고 있습니다. 처음 접해보는 지식이 많아 계속 관련 자료와 웹 사이트를 찾아보며 배우고, 그 내용을 팀원과 협의한 후 프로젝트에 적용시켜 보고 있습니다.

4) 빅데이터 공모전 참가하기 - 내가 공모전에 도전한다면, 선정해 보고 도전, 왜 선정했는지 각자 소감, 각 조별로 의논해서 출전하기

임형열 :
https://m.post.naver.com/viewer/postView.nhn?volumeNo=27869284&memberNo=36383232&vTyp
최근 기존의 농수산업과 IT 기술을 합친 '스마트 팜'과 같은 융복합적 개념이 차세대 핵심기술로 대두되고 있습니다. 또한 우리나라의 농업 규모는 타 OECD 국가대비 매우 높은 수준이며, IT 기술도 세계 최고 수준입니다. 때문에 빅 데이터를 활용하여 무엇인가를 만들어 낸다면, 위에서 언급한 기술과 엄청난 시너지를 낼 수 있을 것이라 생각하여 해당 공모전을 고르게 되었습니다.

김동진 :
https://m.post.naver.com/viewer/postView.nhn?volumeNo=27655737&memberNo=36383232&vType=VERTICAL
이번 프로젝트를 통하여 팀원들과 함께 협업을 하며 데이터분석을 하는 것의 재미를 알게되었습니다. 교내에서 뿐만 아니라 다양한 분야의 사람들과 함께 만나서 빅데이터에 관한 연구를 하고 제가 직접 창업을 할 예정은 아직 없지만 빅데이터를 활용하여 이미 사업을 하는 사람들에게 도움을 주거나 새로운 분석 기술을 생각해 낸다면 빅데이터를 공부한 사람으로서 좋은 기회가 될 것 같아서 이 공모전을 선정했습니다.

조재혁 : https://blog.naver.com/tjddms1022/221910809010
만약 제가 도전한다면, 식품의약품안전처 공공-빅데이터 활용 참업 경진대회를 진행할 것 같습니다. 일반 사용자가 식품의약품의 데이터의 접근성도 어려우므로 가치 있는 데이터를 가공하고 분석하는데 의미가 있는 작업일 것 같고, 제품 서비스를 개발하는 부분에서 작품성도 좋을 것 같아 이 공모전을 선정했습니다.

김태완 : https://kbig.kr/portal/;jsessionid=E5FC412A73AF36113F502D1AD3F5E93D
저는 K-ICT빅데이터 센터에 중소기업 빅데이터 활용지원사업에 참가할 것 같습니다. 일단 현재 다루고 있는 데이터와 비슷하며 이번에 데이터를 가공과 분석을 하며 쌓은 지식을 더 활용하여 도전해 보고 싶습니다.