portfolio

JINHA

Data Analysis & Optimization

Yonsei univ. Oh Jinha

Contents

- I. INTRODUCTION
- II. PROJECT
 - · 1. 2019 START-UP INVESTMENT
 - · 2, 2019 BIG CONTEST
 - · 3. 2020 CONSULTING FIRM
 - · 4. 2020 AI FRIENDS SEASON 1
- III. CODE
- IV. SUMMARY

I . INTRODUCTION

지원자 프로필

지원자는 Python 및 SAS 등 언어를 이용하여 빅데이터 기반 분석을 주로 수행하고 있으며, 20년 2월 연세대학교를 졸업해 경영&전략 컨설팅펌 라이언앤코에서 리서치 및 장표 작성의 업무를 수행하였습니다.

•	생년월일	1995.01.09.
•	연락처	010 4149 57

기본 정보

• 거주지 현재 서울시 관악구 신림동 거주

• 사용프로그램 Python / SAS / Excel & PPT

• 전문분야 Machine Learning & Deep Learning

희망직무 Data Analyst

• **자격증** 데이터분석준전문가(ADsP)

이메일 <u>o41495704@gmail.com</u>

Github https://github.com/lashid/shid

약력				
	2010.03 ~ 2013.02.	전주고등학교 졸업		
	2013.03 ~ 2020.02.	연세대학교 졸업		
0	2019.12 ~ 2020.03.	경영컨설팅펌 라이언앤코 인턴		

* Source:

수행 프로젝트 목록

* Source:

2019년 3월 대학교에서 SAS를 통한 데이터 분석을 배우면서 7월부터는 독학으로 Python을 통한 데이터 분석을 익혔으며, 이를 통해 스타트업 투자 유치 요인 분석, 리니지 유저 잔존가치 예측 등의 프로젝트를 진행하였습니다.

기본 정보						
프로젝트 이름	수행 기간	사용 언어	지원자 수행 역할			
• 스타트업 투자 유치 요인 분석	2019.03. ~ 2019.07.	SAS	로지스틱 및 CBR 분석 및 결과 해석			
• 리니지 유저 잔존가치 예측	2019.07. ~ 2019.09.	Python	LSTM을 통한 시계열 데이터 분석 모델링			
• LG 스마트폰 Feature 선택 방향 제시	2019.10. ~ 2019.12.	Python	설문조사 데이터 기반 Conjoint Analysis			
• 화재 발생 예측	2019.11. ~ 2019.12.	Python	EDA 및 데이터 전처리			
• 부산 AR/VR 산업 중장기 발전전략 수립	2019.12. ~ 2019.12.	-	리서치 및 장표 작성			
• 조직업적평가 지표 개발 및 사업평가 체계 개선	2019.12. ~ 2020.02.	-	리서치 및 장표 작성 & 서울시장 연설 텍스트마이닝			
• 직무분석을 통한 직무중심 보수체계 개선	2020.02. ~ 2020.03.	-	리서치 및 장표 작성 & 평가 툴 제작(Excel)			
• 대전시 센서 온도 예측	2020.03. ~ 2020.04.	Python	EDA 및 LSTM, LGBM 파라미터 튜닝			
		·				

*모든 프로젝트에서 PPT, Excel 사용

II. PROJECT · 1. 2019 START-UP INVESTMENT

로켓펀치에 등록된 정보를 통해, 국내 스타트업이 설립 후 3년 이내에 8억 이상의 투자를 받을 수 있는지 여부를 분석하여 스타트업의 투자 유치 방안 및 투자자들의 전략 제시에 도움을 주고자 하는 것이 본 프로젝트의 목표입니다.

2019 연세대학교 데이터마이닝 이론 및 실습 강의

프로젝트 개요		데이터 세부 내용	
수행 기간	2019.03. ~ 2019.07. (약 4개월)	대표자 출신 학부	범주형 / 서울대, 연세대, 고려대 등 7가지
		대표자 최종 학력	범주형 / 고졸, 대졸, 학사, 석사 등
팀 구성	총 4명 (경영학과, 정보산업공학과)	대표자 팔로워 수	수치형 / (명)
데이터 출처	로켓펀치에 등록된 3년 이상 스타트업 200여개 크롤링	기업의 매력도	수치형 / (팔로우 수를 조회수로 나눈 값)
		기업의 산업 분야	범주형 / 교육, 금융, 패션 등 8가지
목표값	3년 이내에 투자 유치 받은 금액 및 8억 이상의 투자 유치 확률	기업의 기술 분야	범주형 / 웹서비스, 모바일 등 6가지
수집	스타트업 대표 및 기업 자체에 대한 로켓펀치 데이터	기업의 소재지	범주형 / 강남, 강북, 경기, 지방, 기타
데이터		클러스터링 결과	범주형 / Group 1 ~ Group 5

로켓펀치 사이트에서 스타트업 데이터 Crawling을 Python으로 진행하였으며 나머지 데이터 처리 및 모델은 SAS로 구현하였습니다.

분석 기술 스택 크롤링 Python

모델링 S.Sas. SAS

* Source : Python, SAS 홈페이지

기존 선행 연구에서 기업 자체의 수치적인 부분에 집중했던 것과 달리, 대표자의 인프라 및 인맥 요소와 기업의 위치, 매력도 등의 변수를 통해 기존에 없던 새 로운 분류를 하여, 기존과는 다른 접근법을 제시하고자 하였습니다.

FA & Clustering

FA

-0.00160

0.02609 -0.05005

= 5 Average = 1Eigenvalue Difference Proportion Cumulative 0.36638326 0.3137 0,3137 0.5542 1.20224056 0.19960033 0.2404 0,22414519 0.2005 0,7547 0.77849505 0.33049470 0.1557 0.9104 0.44800035 0.0896 1,0000

Rotated Factor Pattern

0,88355 0,05081 -0,02623

0.87195 -0.10336

Factor1 Factor2 Factor3 Factor4

-0.03803 0.99108 -0.10245 -0.02441

-0.00124 -0.10185 0.99454 -0.01535

X4 -0.03843 -0.02389 -0.01513 0.99856

Eigenvalues of the Correlation Matrix: Total

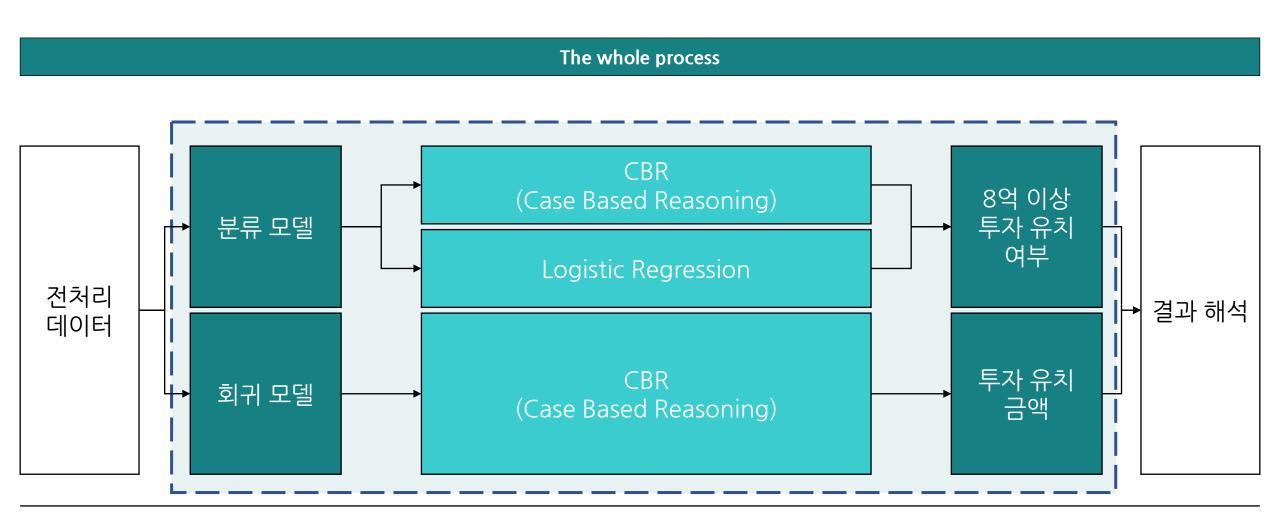
- nfactor=4
- 4번째 성분까지의 누적 설명력 이 91%에 이르므로 요인의 개 수를 4개로 설정
- Factor 해석
- ✔ Factor1 : 대표자 주변의 인프라
- ✓ Factor2 : 대표자의 사교적 성격
- ✓ Factor3: 기업의 위치적 요소
- ✓ Factor4: 기업의 매력도

Clustering

- ncluster=5
- CCC와 Pseudo F Statistics가 크게 증가하고, Pseudo t-Squared가 크게 감소하는 구간을 통해 클러스터의 개수를 5개로 설정
- Cluster 해석
- ✓ Cluster1 : 주변 인프라가 부족한 대표자
- ✓ Cluster2: 서울에 근접한 기업 & 주변 인프라가 좋은 대표자
- ✓ Cluster3 : 서울로부터 먼 기업
- ✓ Cluster4 : 매력도가 높은 기업
- ✓ Cluster5: 대표자가 사교적 성격을 지닌 기업
 - 5개의 Cluster를 새로운 범주형 독립 변수로 추가

* Source : SAS 및 자체 분석 결과

전처리 데이터 중 범주형 변수는 Full-Rank 방식으로 더미화 하였으며, 각각 분류 및 회귀 두 부분으로 나누어서 분석을 진행하였습니다.



결국 로지스틱 회귀 모델을 통해 단순 CBR뿐만 아니라 로지스틱 회귀 모형을 통해 도출된 기업의 투자 유치 확률에 100을 곱하여 0점부터 100점 사이의 수 치화가 가능하며, 투자자 입장에서 기업 간의 비교를 용이하게 하며 데이터 분석 서비스를 이용하게 될 고객에게 맞춤형 서비스를 제공하는 것이 본 모델링의 목표입니다.

제공 서비스 목표

✓ 서비스 제공 프로세스

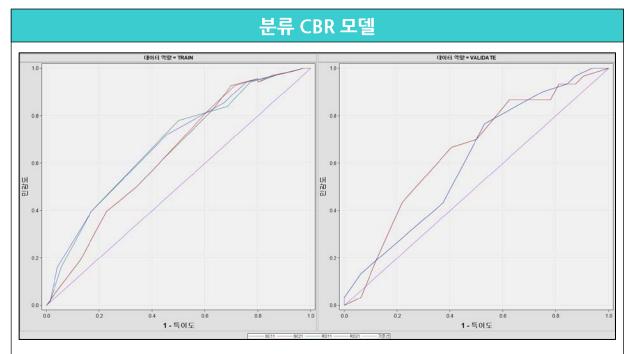
기업 군집화 → CBR을 통해 과거 유사 기업 탐지 → 로지스틱 회귀 확률 점수화

- 먼저 기업 군집화를 통해 이 기업이 갖는 특성을 산업군이 아닌 요인들로 추가 가능
- CBR을 통해 기존 기업들 중 가장 유사한 기업을 찾아 그 기업의 투자 금액과 유치 여부 등을 알 수 있음
- 스타트업의 투자 유치가 기준을 넘을 확률을 로지스틱 회귀를 통해 계산하면 점수화 하여 평가 가능

데이터 분석 결과를 점수화 하여 고객에게 제공 가능

먼저 CBR로, 기존 과거 사례들을 통해 ①8억 이상의 투자 유치 가능성을 분류, ②투자 유치 금액 예측을 실행하였습니다. 그리고 각각 AUC 넓이와 ASE 값을 통해 최적의 Parameter 값을 찾았을 수 있었습니다.

Case-Based Reasoning



분류 모델에서는 SC11(Scan, k=11) 방식이 가장 넓은 AUC를 가진다.

회귀 CBR 모델 선택된 모 선행 노드 타겟 변수 타겟 레이 선택 기 모델 노드 모델 설명 블 준: Valid: Average Squared Error SC21 MBR3 MBR3 X8 39386.64 MBR MBR **RD21** X8 39394.29 X8 40323.38 MBR4 MBR4 SC11 MBR2 MBR2 **RD11** 40323.38

- Validation의 ASE를 기준으로 비교
- SC21 (Scan, k=21) 방식이 가장 작은 ASE를 가짐
- ・ 다만 표준편차가 200(20억원)으로 매우 커서 Outlier 영향 가능성 존재

^{*} Source : SAS 및 자체 분석 결과

로지스틱 회귀 분석을 통해서는 8억 이상의 투자 유치 가능성 만을 분류하였는데, ① 자본 비집약적 산업과 ②자본 집약적 산업을 구분하여 분석하였습니다. 유의미한 데이터인지 여부는 카이제곱 검정을 통해 선택하였으며 그 결과는 아래와 같습니다.

Logistic regression

자본 비집약적 산업

- [유의미한 데이터]
- a. X1-0 : 서울대 / b. X1-1 : 연세대 / c. X1-4 : 인서울 대학 / d. X1-5 : 기 타 국내대학
 - e. X3: 대표자의 인맥 / f. X4: 기업의 팔로우 수 / g. X7-1: 강북 지역 h. X9-1: 주변 인프라 부족한 대표자 / i. X9-2: 서울 근접 및 주변 인프라가 좋은 기업
 - j. X9-3 : 서울로부터 먼 기업
- 총 10개의 유의미한 데이터가 선택되었습니다.
- 이로부터 도출되는 로지스틱 회귀모형은 다음과 같습니다.
- [In[(Odds)=] In](p/(1-p))
 -41.8618x_1.0-35.2940x_1.1-40.2785x._1.4-42.9432x._1.5+0.0157x_3+0.1676x_4+33.1225x_7.1+16.1522x_9.1+14.5077x_9.2+49.5644x_9.3

자본 집약적 산업

- [유의미한 데이터]
- a. X1-6: 해외 대학 / b. X7-1: 강북 지역
- 총 2개의 유의미한 데이터가 선택되었습니다.
- 이로부터 도출되는 로지스틱 회귀모형은 다음과 같습니다.
- $[\ln[(Odds)=]\ln](p/(1-p))$ = -32.2624x_1.6-20.4382x_7.1

* Source : SAS 및 자체 분석 결과

통계적 분석 기법을 배운 후 처음 프로젝트 수행으로, 부족한 면이 많이 존재했으나 가장 많이 고민하고 날을 새는 의미 있는 시간이었습니다. 다만 Outlier에 대한 통제, 절대적인 데이터 수의 부족 등이 개선해야할 점으로 나타났습니다.

개선사항

- 절대적인 **데이터 수 부족**
- 로켓펀치에서 **제공하는 데이터 종류의 한계**
- 투자 유치에 성공한 기업이 아닌, **실패한 기업들에 대한 자료 입수 어려움**
- 투자 유치 성공에 대한 기준 금액 (본 프로젝트에서는 8억 원)을 설정하는 방법이 모호
- 로지스틱 회귀 분석 시 범주별로 매우 큰 차이를 보임
- 투자자와 스타트업 사이의 <u>네트워크에 대한 연구</u>도 이루어지면 좋을 것
- 현재 분석한 기업들의 향후 행보에 따라 실패 기업과 성공 기업 여부가 다시 나누어질 것
- 기업 **군집에 따라** 투지 유치 성공 금액을 **다른 수준**으로 정할 수 있음
- 기본적으로 스타트업 투자금액의 편차가 크기 때문에 <u>Outlier의 영향을 통제</u>하면 더 좋은 결과 예상됨
- 대표자 또는 스타트업 간 관계를 SNA 분석을 통한다면 더 유의미한 변수가 생길 것으로 보임

II. PROJECT
· 2. 2019 BIG CONTEST

리니지 유저의 ①평균 결제 금액, ②게임 지속 기간 예측을 통해 유저의 향후 '잔존가치'를 평가하는 모델을 구축하는 것이 본 프로젝트의 목표입니다.

2019 빅콘테스트 대회 개요 데이터 세부 내용 **LABEL** 유저의 생존 기간과 일별 평균 결제 금액을 제공 대회 기간 2019.07. ~ 2019.09. (약 2개월) **ACTIVITY** 캐릭터별 일일 주요 활동 집계 제공 총 4명 **COMBAT** 캐릭터 전투 활동 정보 일일 집계 제공 팀 구성 (경영학과, 정보산업공학과, 수학과) **PLEDGE** 캐릭터 소속 혈맹 구성원들의 전투 정보 일일 집계 제공 챔피언 리그 - Analysis 분야 **TRADE** 캐릭터 거래 이력(교환, 개인상점) 일일 집계 제공 참가 리그 (대학생 이상 참가 가능) **PAYMENT** 유저 결제 금액 일일 집계 제공 게임 활동 데이터를 활용하여 대회 문제 "게임유저 잔존가치를 고려한 고객 이탈 예측 모형" 개발 제공 리니지 유저의 유저의 **미래 평균 결제 금액 및 게임 지속 기간 도출** 목표 데이터 label, activity, combat, pledge, trade, payment Data

^{*} Source : 2019 빅콘테스트 홈페이지

프로젝트는 기본적으로 Python을 사용하였으며 Google Colab 및 Google Drive로 팀원 간 데이터 및 코드 공유를 하였습니다.

분석 기술 스택

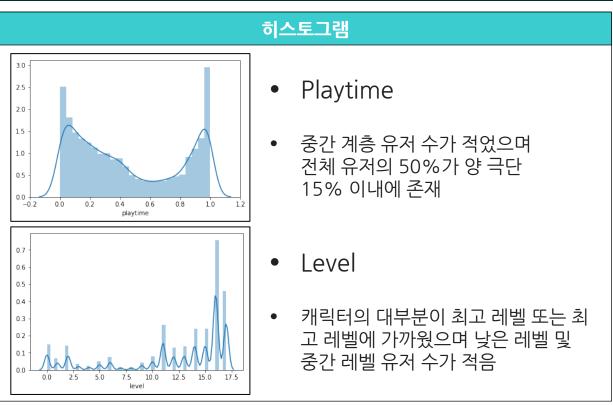


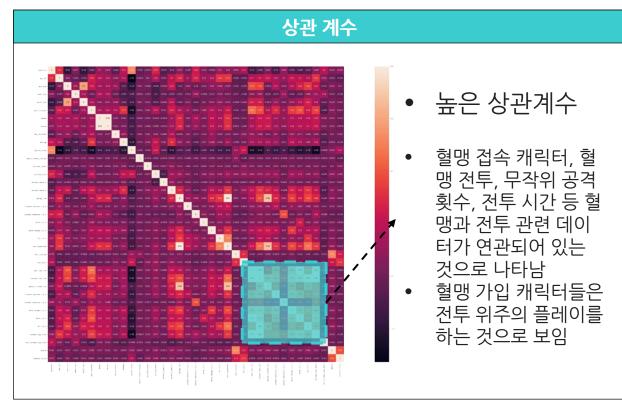


* Source : Python, Google Colab 홈페이지

기본적인 EDA 과정을 통해 리니지 유저의 형태는 다양하며 각각의 형태에 따라 다른 게임 플레이 양상을 보이는 것을 확인하였으며, 이를 통해 전체 유저를 특징 별 군집으로 나눠줘야 할 필요성이 도출되었습니다.

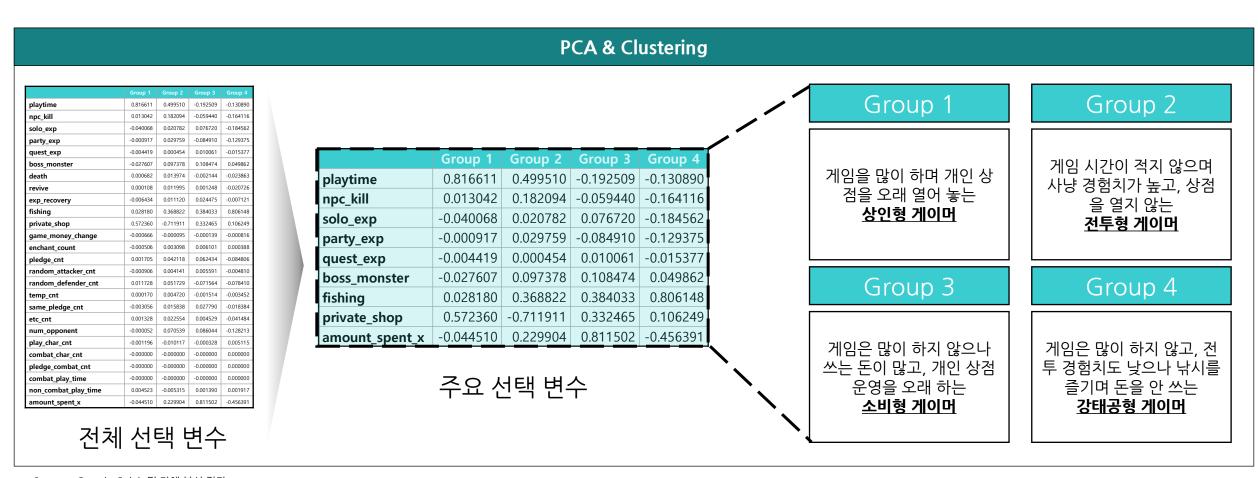
EDA 개요





* Source : Google Colab 및 자체 분석 결과

변수들 중 EDA를 거쳐 선택된 변수들을 통해 전체 데이터에 대해 90.9%의 설명력을 보이는 PCA 4개를 설정하였으며, 이를 통해 유저의 특징을 나눌 수 있 었습니다.



^{*} Source : Google Colab 및 자체 분석 결과

유저의 캐릭터에 따라 구분되어 있는 데이터에서 저희 팀은, '캐릭터'가 아닌 '유저'를 보기 위해 유저 기준으로 데이터를 처리하였으며, 유저 별로 앞서 도출한 PCA 4개의 변수를 추가하였습니다.

데이터 전처리 유저 기준 데이터 처리 유저 A 유저 A 캐릭터 1 1시간 2시간 캐릭터 2 총 4시간 캐릭터 3 1시간 캐릭터 기준 분류 유저 기준 분류

PCA & Clustering

게임을 많이 하며 개인

Group 1

상점을 오래 열어 놓는 상인형 게이머

Group 3

게임은 많이 하지 않으 나 쓰는 돈이 많고, 개 인 상점 운영을 오래 하는 소비형 게이머

Group 2

게임 시간이 적지 않으 며 사냥 경험치가 높고, 상점을 열지 않는 전투형 게이머

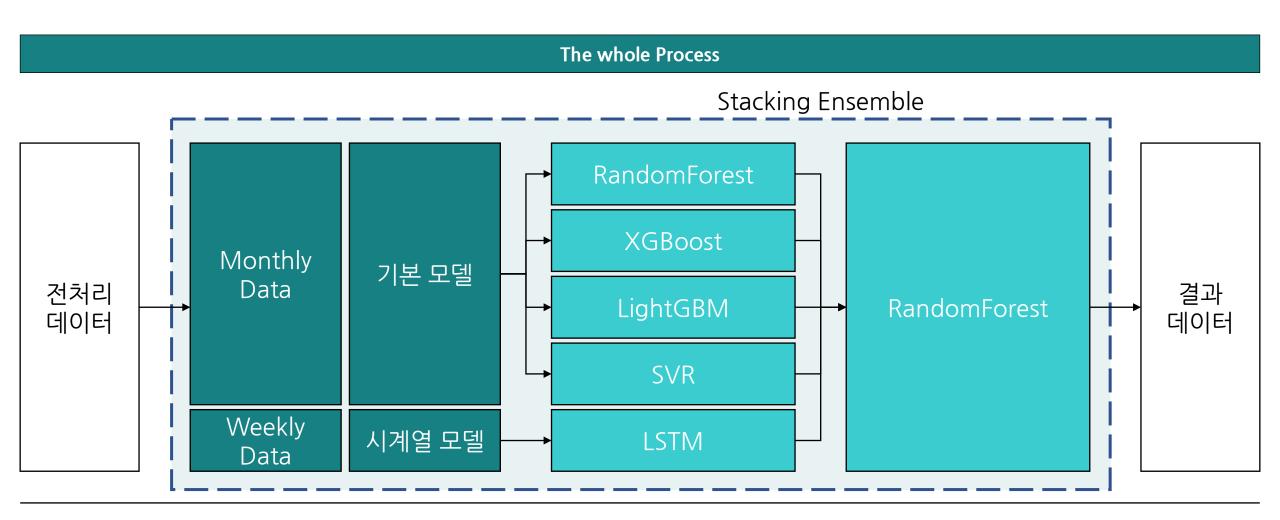
Group 4

게임은 많이 하지 않고, 전투 경험치도 낮으나 낚시를 즐기며 돈을 안 쓰는 강태공형 게이머

- 유저 각각의 성향을 대변할 수 있는 PCA 4개를 변수에 추가함 으로써 기존 데이터가 설명하지 못하는 부분 을 설명할 수 있게 되
- PCA 이외에 다른 변 수를 사용하지 않을 경우, 차원 수를 줄이 는 효과를 가져와 데 이터 처리를 용이하게 할 수 있음

* Source : 자체 분석 결과

Monthly Data로 RandomForest, XGBoost, LightGBM, SVR을 Stacking 하였으며, 시계열 데이터 분석으로 LSTM 을 더하여 분석에 사용하였습니다.



본 지원자가 맡았던 LSTM은 시계열을 반영했음에도 RandomForest, XGB 와 다르지 않은 성능을 보였는데, 이는 데이터 개수가 4만 개로 적었던 것과, 시간 흐름에 따라 유저 흐름이 급변하였기 때문이라는 생각을 하게 되었습니다.

개선사항

- 기존 유저 분류 방식을 개선하여 유저의 실생활 패턴에 따른 분류 (ex. 직장인, 학생 등)
- 복잡하게 쌓은 신경망 모델이 아닌, 오히려 로지스틱 회귀 등 분석을 활용
- 향후 게임 지속 기간과 결제 금액 중, **향후 게임 지속 기간**의 경우, 연속형 변수로 취급하기 보다는, 게임을 계속 한다. OR 게임을 관두었다. 의 <u>Binary 변수</u> 로 취급하였다면 성능이 올라갔을 가능성 존재
- 리니지 게임에 대한 깊은 이해로, 캐시 아이템 및 소비가 어디서 이루어지는지에 대한 정성적인 분석을 더하기

II. PROJECT · 3. 2020 CONSULTING FIRM

라이언앤코의 Data Science Center에서 근무하면서 여러 프로젝트를 진행하였는데, 그 중 스스로 고민하여 개발자 적인 생각을 한 프로젝트를 소개하려 합니다. 조직업적평가 지표 개발에서는 '서울디자인재단의 성과'가 무엇인지 고민해야 했는데, 서울시 시장의 연설문을 텍스트 마이닝하여 주요 방향성을 제시해 주었습니다. 또한 직무분석에서는 엑셀 함수를 활용하여 직무 분석 툴을 만들 수 있었습니다.

2020 라이언앤코 인턴

조직업적평가 지표 개발 및 사업평가 체계 개선

Seoul Design 서울디자인재단

- KoNLPy 를 통해 한국어 불용어 제거 후
- Embedding: TF-IDF
- 그 후 중요도 분석을 통해
 시민, 청년, 신혼, 창업, 돌봄이라는 키워드 도출

직무분석을 통한 직무중심 보수체계 개선



- 평가 위원들이 직무 별 중요도를 입력하면 가중치를 곱해 각 직무의 중요도를 도출하고, 그래프로 보여주는 엑셀 파일 작성
- 파일은 실제로 고객사에 제공되었음

* Source: 각 사 홈페이지, 실제 분석 자료는 고객사에게만 제공하게 되어 있어 설명으로 대체

II. PROJECT

· 4. 2020 AI FRIENDS SEASON 1

33일의 연속된 기상청의 8개 변수 공공데이터 및 30일의 비교 센서 18개 온도, 그 후 3일의 목표 센서 1개 온도를 통해 그 후 80일 동안의 목표 센서 온도를 예측하는 것이 본 프로젝트의 목표입니다.

2020 Al Friends Season 1 대회 개요 데이터 세부 내용 대회 기간 2020.03. ~ 2020.04. (약 1개월) 기온, 현지기압, 풍속 등 8개 변수 데이터 30일 총 2명 팀 구성 (정보산업공학과, 수학과) 센서 18개의 온도 데이터 대회 주최 AI프렌즈, 한국원자력연구원, 한국기계연구원, DACON 기온, 현지기압, 풍속 등 8개 변수 데이터 3일 목표 센서(y) 1개의 온도 데이터 대회 문제 기상청 공공데이터를 활용한 대전시 온도 추정 제공 대전시 특정 지역 센서의 30+3일 간의 그 후 80일의 목표 센서(y) 온도 예측 데이터 기온, 현지기압, 풍속, 일일 누적강수량, 해면기압, 습도 등

* Source : 데이콘 홈페이지

프로젝트는 기본적으로 Python을 사용하였으며 Google Colab 및 Google Drive로 팀원 간 데이터 및 코드 공유를 하였습니다.

분석 기술 스택

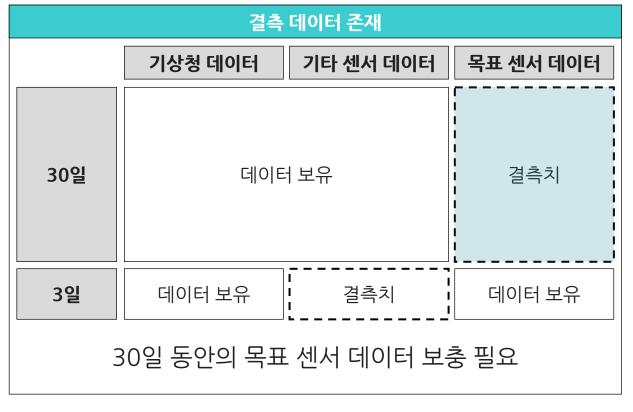


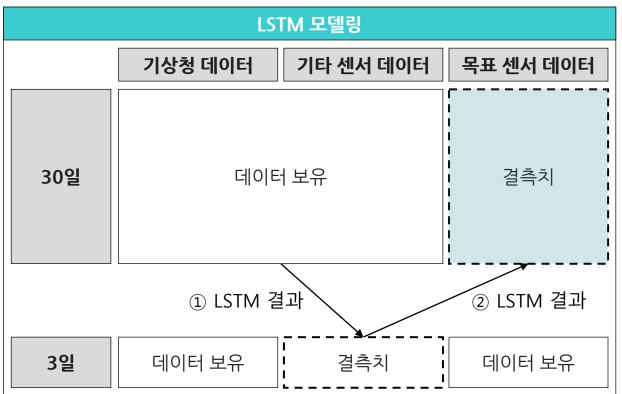


* Source : Python, Google Colab 홈페이지

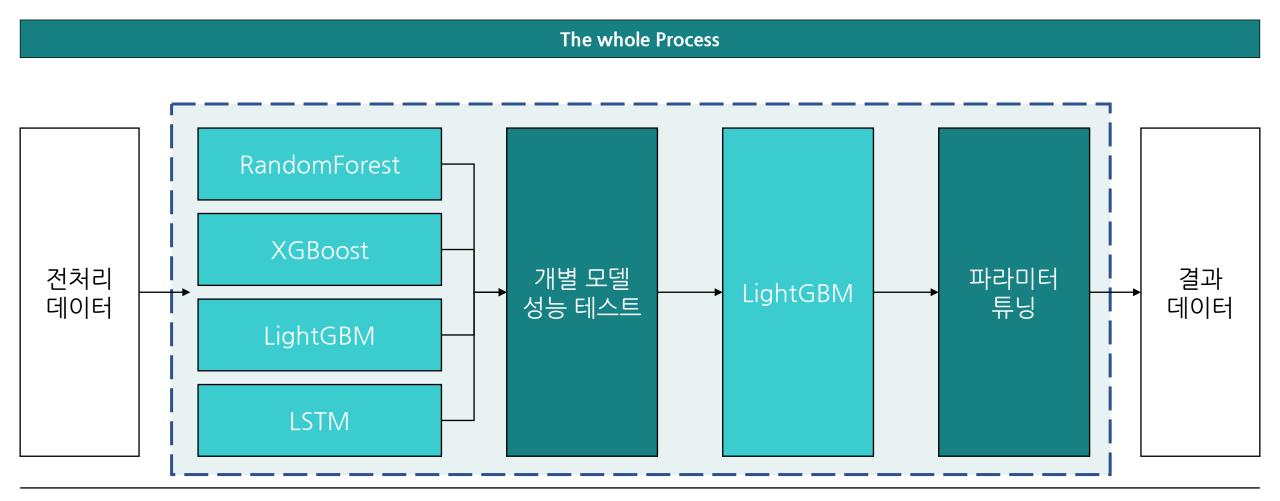
주어진 센서는 총 19개로 그 중 18개의 센서 데이터는 30일치가 주어졌으나 예측해야 하는 1개의 센서 데이터는 3일치만 주어졌습니다. 그래서 학습 데이터의 결측치를 채우는 것이 중요했고, 저희는 LSTM의 결과 값을 통해 이를 해결하여 총 33일의 학습 데이터를 얻을 수 있었습니다.

결측치 대체 과정





전처리 데이터를 바탕으로 Xgb, Random Forest, Lgbm, LSTM 등 모델의 테스트 결과 가장 성능이 좋았던 Lgbm의 Parameter를 튜닝하는 것으로 방향을 잡았습니다.

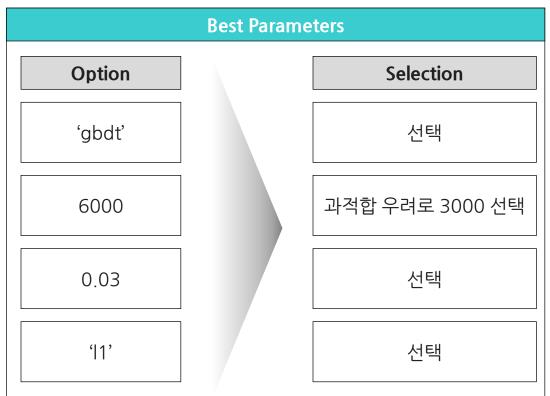


4. 대전시 센서 온도 예측 - Modeling (2/2)

Ⅱ. 수행 프로젝트

LightGBM의 Parameter 중 boosting_type, n_estimators, learning_rate, metric을 중요하게 튜닝하였으며 이 과정에서 GridSearchCV를 활용하였습니다.

Parameter Tuning LGBM Parameter Tuning Parameter Range boosting_type 'gbdt', 'dart' 1000, 2000 ··· 6000 n estimators **Parameter Tuning** learning_rate 0.01, 0.02 ··· 0.1 'l1', 'l2', 'rmse', 'mape', metric 'gamma'



최종 성적은 379팀 중 81등이었으며 가채점보다 최종 순위가 80등 가량 높았습니다. 과적합을 우려했기에 Ensemble 등 모델을 쌓지 않았기 때문이라 생각합니다. 다만, 초반 EDA에서 결측치를 채우는 과정에서 실내와 실외 등 센서들의 특성을 구분하였다면 더 좋은 결과가 나왔을 거라 생각합니다.

개선사항

- 제시된 18개 센서 데이터의 그룹을 나누어 목표 센서 매핑
- 성능이 가장 좋았던 모델 하나만 쓰는 게 아니라, Ensemble Stacking 등 기법으로 더 깊게 학습 및 여러 모델의 장점을 활용할 필요가 있음
- 결측 데이터를 채울 때 LSTM의 시간 주기 설정 및 은닉 층 조정
- 목표 데이터가 센서의 '온도'이기 때문에 기상청 데이터에서 기온이 가장 중요했을 것이며 이를 활용할 방안 필요

III. CODE

위 프로젝트 뿐만 아니라 지원자 Github의 Project 폴더에 다른 프로젝트에서 활용했던 언어 및 분석 코드가 제시되어 있습니다. ☺

지원자 Github 주소

https://github.com/lashid/shid

IV. SUMMARY

마무리 멘트

전공에서 배워 온 ①통계 기반 정량적 분석과, 경영 수업 및 컨설팅 펌 인턴을 통해 배워 온 ②논리적이고 직관적인 정성적 분석이 더해져, 현재 지원자가 되었습니다. 지금부터 10년 간은 바보라고 불릴 정도로 일만 해 볼 생각이며, 머리가 제일 잘 돌아가는 시기에 머물 회사를 찾고 있습니다. 잘 부탁드리겠습니다. 감사합니다.

지원자의 강점 분류

정량적 분석 강점 SAS 통계 수학 머신러닝 Python 딥러닝

정성적 분석 강점 경영 컨설팅 사업 관리 소규모 스타트업 1년 근무 감사합니다!

지원자 오진하 드림