portfolio

JINHA

Data Analysis & Optimization

Yonsei univ.
Oh Jinha

Contents

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.

기본 정보

- **생년월일** 1995.01.09.
- **연락처** 010.4149.5704
- **거주지** 현재 서울시 관악구 신림동 거주
- **사용프로그램** Python / SAS / Excel & PPT
- **전문분야** Machine Learning & Deep Learning
- **희망직무** Data Analyst
- **자격증** 데이터분석준전문가(ADsP)

- **이메일** o41495704@gmail.com
- **Github** https://github.com/lashid/shid

약력

-------○	2010.03 ~ 2013.02.	전주고등학교 졸업
-------○	2013.03 ~ 2020.02.	연세대학교 졸업
-------○	2019.12 ~ 2020.03.	경영컨설팅펌 라이언앤코 인턴

* Source :

수행 프로젝트 목록

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 모델 개발**	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)
• **(진행 중) 대전시 센서 온도 예측 공모전**	2020.03. ~ 현재	Python	EDA 및 LSTM의 Layer 구축 진행 중

* Source :

*모든 프로젝트에서 PPT, Excel 사용

Ⅱ. PROJECT

· 1. 2019 START-UP INVESTMENT

1. 스타트업 투자 유치 요인 분석 - 개요 (1/3)

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

수행 기간	2019.03. ~ 2019.07. **(약 4개월)**
팀 구성	총 4명 (경영학과, 정보산업공학과)
데이터 출처	로켓펀치에 등록된 3년 이상 스타트업 200여개 크롤링
목표값	3년 이내에 투자 유치 받은 금액 및 8억 이상의 투자 유치 확률
수집 데이터	스타트업 대표 및 기업 자체에 대한 로켓펀치 데이터

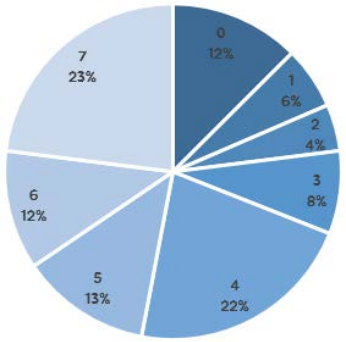
데이터 세부 내용

대표자 출신 학부	범주형 / 서울대, 연세대, 고려대 등 7가지
대표자 최종 학력	범주형 / 고졸, 대졸, 학사, 석사 등
대표자 팔로워 수	수치형 / (명)
기업의 매력도	수치형 / (팔로우 수를 조회수로 나눈 값)
기업의 산업 분야	범주형 / 교육, 금융, 패션 등 8가지
기업의 기술 분야	범주형 / 웹서비스, 모바일 등 6가지
기업의 소재지	범주형 / 강남, 강북, 경기, 지방, 기타
클러스터링 결과	범주형 / Group 1 ~ Group 5

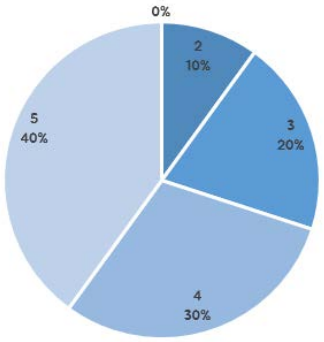
1. 스타트업 투자 유치 요인 분석 - 개요 (2/3)

변수 기초 통계량

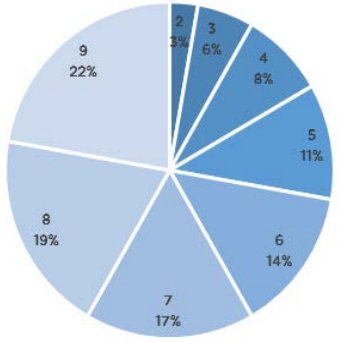
범주형 변수



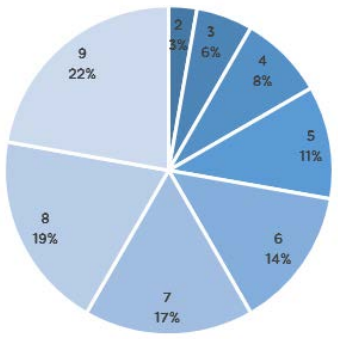
대표자 출신 학부



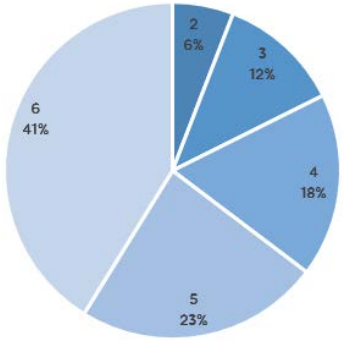
대표자 최종 학력



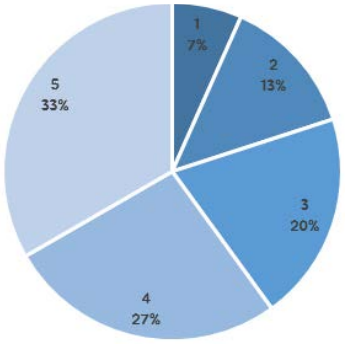
기업의 산업 분야



기업의 기술 분야



기업의 소재지



클러스터 결과

연속형 변수

변수	개수	평균	표준편차
대표자 팔로워 수	200.00	133.50	236.24
기업의 매력도	200.00	14.79	17.44
투자유치 금액(Y)	200.00	180.98	279.46

* Source : 로켓펀치 홈페이지 및 자체 분석 결과

1. 스타트업 투자 유치 요인 분석 - 개요 (3/3)

전체 프로세스

✓ 기업 분석 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**

스타트업의 특성을 새로운 변수로 정의 가능
대표자 학력, 산업군, 매력도 등의 정성적인 자료를 수치화 가능

1. 스타트업 투자 유치 요인 분석 - EDA (1/1)

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

FA & Clustering

FA

Eigenvalues of the Correlation Matrix: Total = 5 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	1.56862381	0.36638326	0.3137	0.3137
2	1.20224056	0.19960033	0.2404	0.5542
3	1.00264023	0.22414519	0.2005	0.7547
4	0.77849505	0.33049470	0.1557	0.9104
5	0.44800035		0.0896	1.0000

Rotated Factor Pattern

	Factor1	Factor2	Factor3	Factor4
X2	0.88355	0.05081	-0.02623	-0.00160
X1	0.87195	-0.10336	0.02609	-0.05005
X3	-0.03803	0.99108	-0.10245	-0.02441
X7	-0.00124	-0.10185	0.99454	-0.01535
X4	-0.03843	-0.02389	-0.01513	0.99856

- nfactor=4
- 4번째 성분까지의 누적 설명력이 91%에 이르므로 요인의 개수를 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

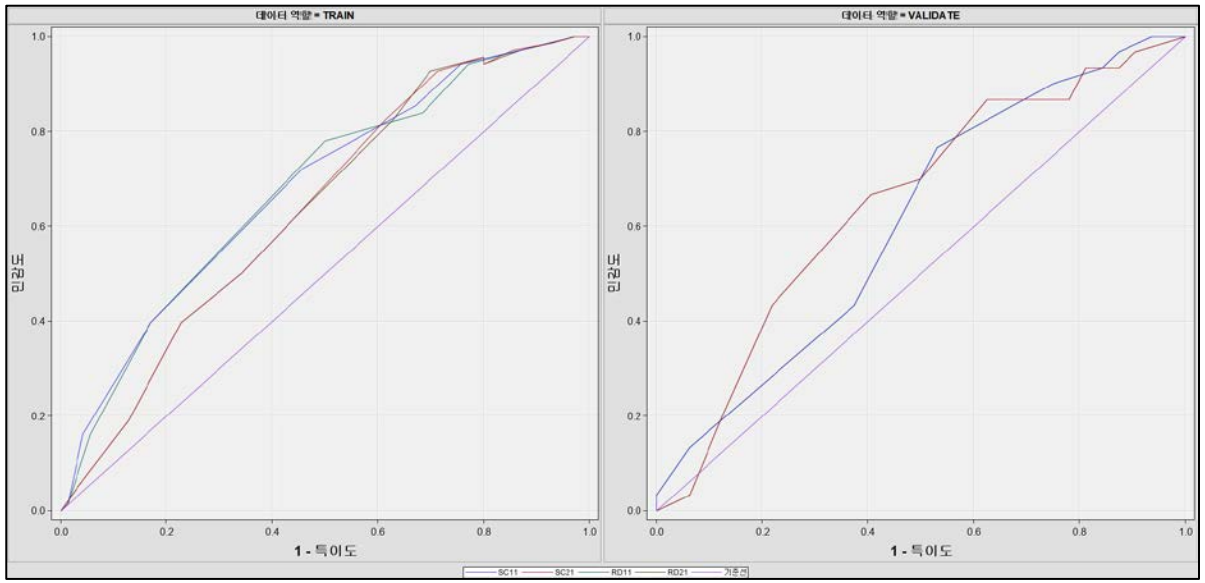
* Source : SAS 및 자체 분석 결과

1. 스타트업 투자 유치 요인 분석 - Modeling (1/2)

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 CBR 모델



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

회귀 CBR 모델

선택된 모델	선행 노드	모델 노드	모델 설명	타겟 변수	타겟 레이블	선택 기준: Valid: Average Squared Error
Y	MBR3	MBR3	SC21	X8		39386.64
	MBR	MBR	RD21	X8		39394.29
	MBR4	MBR4	SC11	X8		40323.38
	MBR2	MBR2	RD11	X8		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

1. 스타트업 투자 유치 요인 분석 - Modeling (2/2)

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 : 기타 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.**

- $\ln(Odds) = \ln(p/(1-p))$
  = -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

1. 스타트업 투자 유치 요인 분석 - 개선사항

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 **네트워크에 대한 연구**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 **Outlier의 영향을 통제**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

2. 리니지 유저 잔존가치 예측 - 개요

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

대회 기간	2019.07. ~ 2019.09. **(약 2개월)**
팀 구성	총 4명 (경영학과, 정보산업공학과, 수학과)
참가 리그	챔피언 리그 - Analysis 분야 (대학생 이상 참가 가능)
대회 문제	게임 활동 데이터를 활용하여 **"게임유저 잔존가치를 고려한 고객 이탈 예측 모형" 개발**
제공 데이터	리니지 유저의 label, activity, combat, pledge, trade, payment Data

데이터 세부 내용

LABEL	유저의 생존 기간과 일별 평균 결제 금액을 제공
ACTIVITY	캐릭터별 일일 주요 활동 집계 제공
COMBAT	캐릭터 전투 활동 정보 일일 집계 제공
PLEDGE	캐릭터 소속 혈맹 구성원들의 전투 정보 일일 집계 제공
TRADE	캐릭터 거래 이력(교환, 개인상점) 일일 집계 제공
PAYMENT	유저 결제 금액 일일 집계 제공

유저의 **미래 평균 결제 금액 및 게임 지속 기간 도출** 목표

* Source : 2019 빅콘테스트 홈페이지

2. 리니지 유저 잔존가치 예측 - 기술 스택

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

분석 기술 스택

언어

Python

공유 플랫폼

Colab

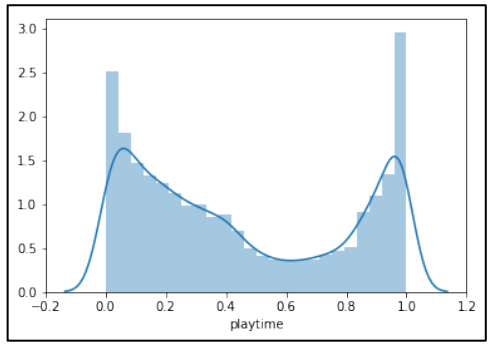
* Source : Python, Google Colab 홈페이지

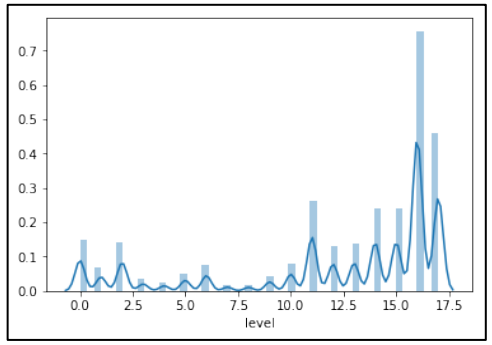
2. 리니지 유저 잔존가치 예측 - EDA (1/3)

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

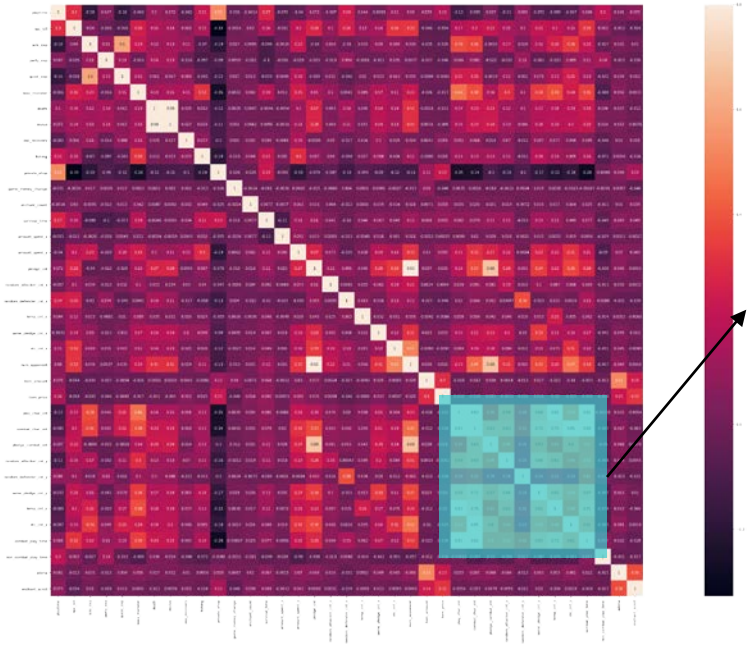
히스토그램





- Playtime
- 중간 계층 유저 수가 적었으며 전체 유저의 50%가 양 극단 15% 이내에 존재
- Level
- 캐릭터의 대부분이 최고 레벨 또는 최고 레벨에 가까웠으며 낮은 레벨 및 중간 레벨 유저 수가 적음

상관 계수



- 높은 상관계수
- 혈맹 접속 캐릭터, 혈맹 전투, 무작위 공격 횟수, 전투 시간 등 혈맹과 전투 관련 데이터가 연관되어 있는 것으로 나타남
- 혈맹 가입 캐릭터들은 전투 위주의 플레이를 하는 것으로 보임

* Source : Google Colab 및 자체 분석 결과

2. 리니지 유저 잔존가치 예측 - EDA (2/3)

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

PCA & Clustering

	Group 1	Group 2	Group 3	Group 4
playtime	0.816611	0.499510	-0.192509	-0.130890
npc_kill	0.013042	0.182094	-0.059440	-0.164116
solo_exp	-0.040068	0.020782	0.076720	-0.184562
party_exp	-0.000917	0.029759	-0.084910	-0.129375
quest_exp	-0.004419	0.000454	0.010061	-0.015377
boss_monster	-0.027607	0.097378	0.108474	0.049862
death	0.000682	0.013974	-0.002144	-0.023863
revive	0.000108	0.011995	0.001248	-0.020726
exp_recovery	-0.006434	0.011120	0.024475	-0.007121
fishing	0.028180	0.368822	0.384033	0.806148
private_shop	0.572360	-0.711911	0.332465	0.106249
game_money_change	-0.000666	-0.000095	-0.000139	-0.000816
enchant_count	-0.000506	0.003098	0.006101	0.000388
pledge_cnt	0.001705	0.042118	0.062434	-0.084806
random_attacker_cnt	-0.000906	0.004141	0.005591	-0.004810
random_defender_cnt	0.011728	0.051729	-0.071564	-0.078410
temp_cnt	0.000170	0.004720	-0.001514	-0.003452
same_pledge_cnt	-0.003056	0.015838	0.027790	-0.018384
etc_cnt	0.001328	0.022554	0.004529	-0.041484
num_opponent	-0.000052	0.070539	0.086044	-0.128213
play_char_cnt	-0.001196	-0.010117	-0.000328	0.005115
combat_char_cnt	-0.000000	-0.000000	-0.000000	0.000000
pledge_combat_cnt	-0.000000	-0.000000	-0.000000	0.000000
combat_play_time	-0.000000	-0.000000	-0.000000	0.000000
non_combat_play_time	0.004523	-0.005315	0.001390	0.001917
amount_spent_x	-0.044510	0.229904	0.811502	-0.456391

전체 선택 변수

	Group 1	Group 2	Group 3	Group 4
playtime	0.816611	0.499510	-0.192509	-0.130890
npc_kill	0.013042	0.182094	-0.059440	-0.164116
solo_exp	-0.040068	0.020782	0.076720	-0.184562
party_exp	-0.000917	0.029759	-0.084910	-0.129375
quest_exp	-0.004419	0.000454	0.010061	-0.015377
boss_monster	-0.027607	0.097378	0.108474	0.049862
fishing	0.028180	0.368822	0.384033	0.806148
private_shop	0.572360	-0.711911	0.332465	0.106249
amount_spent_x	-0.044510	0.229904	0.811502	-0.456391

주요 선택 변수

Group 1

게임을 많이 하며 개인 상점을 오래 열어 놓는 **상인형 게이머**

Group 2

게임 시간이 적지 않으며 사냥 경험치가 높고, 상점을 열지 않는 **전투형 게이머**

Group 3

게임은 많이 하지 않으나 쓰는 돈이 많고, 개인 상점 운영을 오래 하는 **소비형 게이머**

Group 4

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

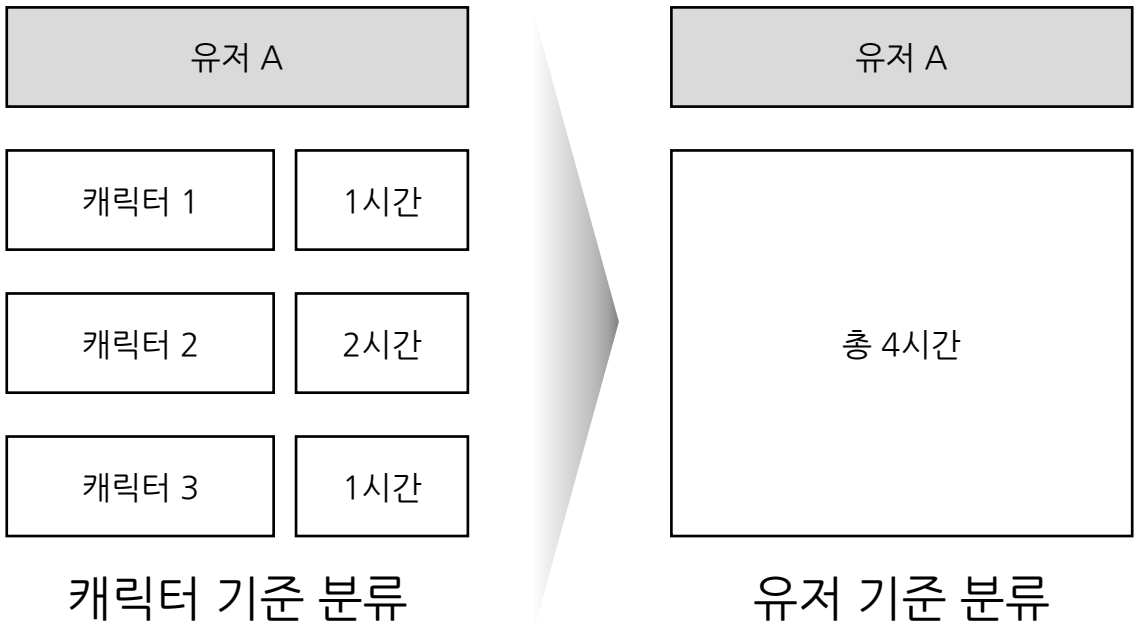
* Source : Google Colab 및 자체 분석 결과

2. 리니지 유저 잔존가치 예측 - EDA (3/3)

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



PCA & Clustering

Group 1

게임을 많이 하며 개인 상점을 오래 열어 놓는 **상인형 게이머**

Group 2

게임 시간이 적지 않으며 사냥 경험치가 높고, 상점을 열지 않는 **전투형 게이머**

Group 3

게임은 많이 하지 않으나 쓰는 돈이 많고, 개인 상점 운영을 오래 하는 **소비형 게이머**

Group 4

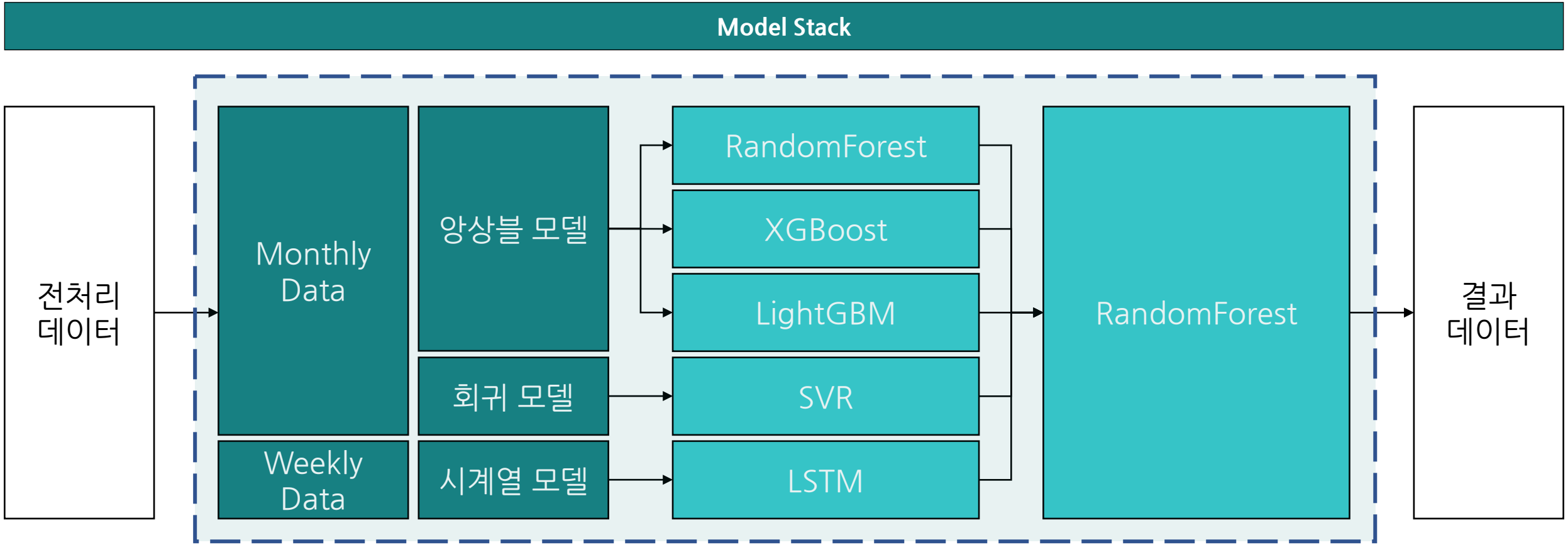
게임은 많이 하지 않고, 전투 경험치도 낮으나 낚시를 즐기며 돈을 안 쓰는 **강태공형 게이머**

- 유저 각각의 성향을 대변할 수 있는 PCA 4개를 변수에 추가함으로써 기존 데이터가 설명하지 못하는 부분을 설명할 수 있게 되었음
- PCA 이외에 다른 변수를 사용하지 않을 경우, 차원 수를 줄이는 효과를 가져와 데이터 처리를 용이하게 할 수 있음

* Source : 자체 분석 결과

2. 리니지 유저 잔존가치 예측 - Modeling

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



2. 리니지 유저 잔존가치 예측 - 개선사항

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 **Binary 변수로 취급**하였다면 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

3. 컨설팅펌 프로젝트

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

울산항만공사 ULSAN PORT AUTHORITY

- 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

Ⅲ. CODE

지원자 코드

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

통계

SAS

수학

Python

머신러닝

딥러닝

정성적 분석 강점

경영

컨설팅

사업 관리

소규모 스타트업 1년 근무

감사합니다!

지원자 오진하 드림