프로젝트 발표: 중간발표는 20%

나머지 2명에서

* 팀 설명

\* Engineering

각종 기업들의 채용공고들

커리큘렴에서 배운다는 것을 강조

현업이랑 가장 많이 연결이 된다.

선배: scientist를 구할 때 spark를 쓸 수 있는 사람들만 뽑는다.

\* Design

데이터를 설계하는 의미의 디자인이다.

적용위주로 해볼 사람들

스펙관리를 위주로 시각화, ML 외에 스터디를 할 수 있다.

\* Scentist

이론, 알고리즘을 배우는 것을 중점으로 하고 있다.

ML / DL 팀을 나눠서 이루어진다. 완전 각자 다른 트랙으로 갈 예정

DL: 논문을 많이 보고, Stanford 강의를 많이 본다.

ML: 책을 중심으로 많이 공부. 매주 토요일에 발제

DL 학기 중에 논문을 구현해보는 것을 하려고 하고 있다

기초가 부족하고 Data science를 알고 싶은 사람은 science팀에 들어와라

\* 교육팀

학기 중에는 '지식채널 Y'를 통해서 지식을 전한다.

방학 때는 교육부장이 신입생 교육을 위해서 계획을 하는 식임

일이 많이 없으니까 많이 들어와달라.

\* 기획부

하는 게 MT 기획, 학기 중에는 전체세션에서 놀러가는 콘텐츠

체육대회: 병맛 콘텐츠

더 지니어스

신뢰에서 머리 쓰면서 머리 쓰면서 하는 것.

들어오면 학기 중에 준비하고 MT 한 번 준비하고 그 다음 학기 행사를 진행할 것이다.

나만 아니면 된다하면 기획부가 좋다.

\* PR부

부장 취임한지 2일.

PR부는 홍보가 주 목적

페이스북 관리. 교육세션 마무리해서 올리고

1달에 1번 만들어서 페이스북에 올림

콘프런스 홍보 포스터도 만든다.

1학기에 1번 정도 일하게 된다.

<머신러닝 1조 발표>

목표, 둘러보기, 데이터분석(모델), 추후계획

transaction row, identity row 설명. 25%만 identity 정보를 가지고 있다.

identity의 모든 행이 채워져 있다해도 merge하면

결측치가 75%

각 변수설명

TransactionAmt: log 스케일로 바꿈

id\_30:

id\_31: browser 별로 label

DeviceType:

\* V-column

339개. 자의적 판단이 어려워서 직관적으로 하는 EDA는 힘들다고 생각함. RFECV를 사용. scikilearn에 있는 method.

LGBMClassifier를 estimator로 선정

softvector classifier

\* Transaction DT

시간 뽑아냄

\* emaildomain

하나의 열로 합치고 삭제

\* card, addr

value count로 해서 봄

숫자가 category를 알아내는 것을 볼 수가 있었음

count값을 feature값으로 새로 활용하는 방법을 차용함

\* XGBoost 모델을 봄

이전에 regression은 50% 정도 나와서 ensemble을 사용하기로 함

94% ROC score가 나옴

Extreme Gradient Boosting

XGBoost가 빨리 병렬처리를 해줌

결측치도 처리해준다.

Kaggle 대회에서 애용되고 있다.

\* 앞으로 계획

전처리 feature중에서 RFE를 돌리고

Raw data를 가지고 처음으로 모든 변수로 RFE를 할 예정

새로운 모델 활용: lightGBM ,KNeightborClassifier를 활용할 계획

<머신러닝 2조 발표>

데이터 셋 소개 -> 전처리 및 EDA & FE -> 데이터 통합 -> 모델 결과 -> 추후 계획

전처리과정

각 column 별 전처리

\* transaction Data

undersampling을 통해서 해결하기 위해서 노력함

\* TransacitonDT

hour column

5시 ~ 10시까지 사기거래 비중이 높다.

\* 요일

요일별 차이가 적다.

\* Amt

거래금액 중 30,000 이상 금액을 없앴음

\* ProductCD

one-hot encoding

\* Card 1~6

먼저 자료의 형태를 바꿔주었다.

\* Card1, 2, 5

범주형이기 하지만 범주의 개수가 많아서 더미화 x.

각 카테고리별 비율값으로 채워주고 원래 컬럼은 지웠다.

\* Card3, 4, 6

공통적으로 one-hot encoding을 했음

card3: 150, 185, etc 3개 category 변수로 바꿈

card6: 제거하고 one-hot encoding을 했음

\* addr1, addr2

addr2만 이용하기로 결정한 뒤에 더미변수화 처리했다

\* emaildomain

R\_email은 사용 x

P-email

\* C-columns

어떤 것을 나타내는지 힘들다.

개인 정보 보호를 위해서 mask했다.

일련의 번호를 만들어서 칼럼을 생성

일련번호 비율을 계산해서 사용함

\* D-columns

Time\_delta, numeric data

column의 결측치는 최빈값인 0으로 대체

binary column을 생성

\* M-column

기재된 내용이 사실인지 여부를 나타내는 columns

결측치를 missing이라는 값으로 채워주었다.

더미화 해서 상관계수가 0.75이상일 경우에는 칼럼 하나만 남기고 drop

결과 17개의 column만 남았다.

\* V-column

랭킹, 집계를 나타내는 columns

결측치 비율이 47% 이상인 column drop

상관계수가 0.75 이상일 경우에는 하나만 남기고 drop

\* Identity

연속형 or 범주형을 나눔

결측치 벼율에 따라서 6% 미만일 때는 결측치 row 제거

10~11%, 50~70% 결측치가 있으면 관측치들이 고르게 분포했기에 표본 값을 random하기 imputation

결측치 비율 96% 이상은 column을 그냥 삭제했음

\* id\_12, 15

결측치 row 제거 후에 dummy 화를 시켰음

\* id\_28, 29

두 column을 비교하고 하나만 남기고 일치 여부를 나타내는 더미변수를 추가한 뒤에 제거

\* id\_30, id\_33

결측치가 충분히 많기 때문에 drop

\* id\_31, 32, 34

더미화

\* Deviceinfo, DeviceType

info는 버리고 type은 더미화 시켰다.

\* 1) transaction & identity

2) Only transaction

이렇게 데이터를 나눠서 예측을 진행하였다.

\* 모델: Gradient Boost을 사용

Randomundersampling을 실시!

최적 parameter를 찾기 위해 그리드 search를 하였을 때

f1\_score가 84%

전체 transaction data를 돌려본 결과 59%가 나왔다. 더 높이는 것이 목표

\* 계획

tansaction + identity 데이터 셋도 모델 돌려보기

Imbalance 문제를 해결하기 위해 oversampling을 사용

다른 모델을 돌려보기

학습 속도를 높이기

Precision/Recall trade-off 고려하면서 모델링하기

c-column은 null값이 없어서 특이하다. c1~c14까지를 groupby를 했을 때 특정 group내에서 데이터가 뭉쳐져 있었다. 그걸 일련번호를 두었을 때도 의미가 있다고 생각해서 그렇게 했다.

Q. 왜 undersampling으로 했는지?

불균형한 데이터를 처리하는 방법 중 하나를 해보고자 했음

<딥러닝 팀>

GAN

정학이와 아이들. 아이 1번 조준흠

목차: GAN -> DCGAN -> KAGGLE -> CODE -> 최종발표방향

\* GAN

가짜로 만든 얼굴들

\* DCGAN

GAN의 불안정한 학습을 개선시킨 것

<머신러닝 4조>

모델링, 빠르고, 정확하게, 참/거짓

TransactionDT: Test와 Train data의 overlapping area가 없다.

Amt에 로그를 씌우기!

ProductCD를 중요하다고 생각

Card4: 비자, 마스터 카드

card6: debit, credit이 제일 높다

Addr1: 도시 주소

Addr2: 국가 주소

dist1, dist2: 거리값. 결측치가 많아서 column을 drop

Pemaildomain: 구매자 이메일 도메인

Remaildomain: 수신자 이메일 도메인

C, D, V: 동일 비율 결측치 변수들을 하나의 대표 column만 골라서 나머지를 drop시킴. 대포 컬럼 중에서도 80% 결측치가 있으면 drop했다.

identity

\* 결측치 처리

방법소개

결측치비율

결측치처리

addr1: 도시

addr2: 국가

하나도 빠짐없이 일치하였기에 통합했다.

결측치가 있는 값들이 동일

drop하고 통합해서 줄임

card4, card6: 범주가 많지 않아서 대체 가능하다고 생각함

Deviceinfo: unknown으로 처리!

모델: Decision Tree

\* addr2 & P\_emaildomain

P\_emaildomain에서 국가를 추출했다.

co.uk / co.france. 등등으로 국가를 추출할 수 있었다.

addr2은 87번이 미국이다.

UK, Germany, France -> 30, 32 각각을 국가와 유의미한 관계가 있을 것이라고 생각함

\* DeviceInfo

cramer's V를 사용함

결측치가 90%가 넘어가서 쓸모없다고 판단해서 DeviceInfo만 쓰기로 함

정리해보면

이용변수를 몇 개로 추렸음.

<건욱이 comment>

데이터를 압축하는 방법이 있다.

학습에 오래 걸리니까 압축을 하는 방법이 있다!

<차후 계획>

모델 안에서 parameter를 조절해서 f1\_score 올리기

점수가 높게 나온 거는 표본을 뽑은 것이다. -> 의미가 없다.

columns가 줄이던가, Null값 처리할 때 binary로 처리했는데 그렇게 하지 말고 고유값들을 0, 1처리하고 null값은 2로 처리해서 해보기. row를 뺄지말지도 정해야 할 것 같다.