2019.08.13 팀 프로젝트 미팅

시작시작시작시작

Memory

**# Device-info**

Device-info 바꿔보려고 했는데 1,700개가 넘는다.

그러면 '제조사'로 묶는 게 쉽지 않을까?

이용하면 좋을 것 같기는 한데 분포가 크고 1개짜리가 너무 많다.

상관관계 구할 때 Pearson 상관관계를 사용합니다.

**# id\_31**

**# Tree model**

너무 잘 나와서 문제

profiling에서 상관관계 있는 것들 다 가져오고

null값들은 다 0으로 채워서 Tree를 하니까 97%가 나온다.

Kaggle은 F1 score다.

Tree model은 그냥 accuracy가 97%가 나온다.

지금 학습이 아예 안 된거다.

Validation을 잘못 나눠서

Imbalance를 고려하지 못했다.

무조건 Fraud에 맞추면 데이터가 적어지니까 down-sampling, over-sampling을 해야 할 필요가 있을 것 같다.

Validation 뽑을 때도 비율 맞춰서 k-fold로 뽑아야 할 것 같다.

**# card3**

addr2가 가장 많은 거 3개가 시간 분포가 똑같다.

미국, 캐나다, 멕시코 일 것 같다.

card3, card5를 미미하게 correlation이 있다.

card5는 fraud transaction 비율대로 section을 구분하니까 corr가 나왔다

card1,2,4,6은 correlation이 거의 안 나왔다.

**# V-column**

correlation

isfraud와의 corr이 0.36, 0.38이 나오기는 했다.

**# repeated transaction**

연관성을 많이 찾지는 못하였는데, fraud거래에서의 repeated transaction이 fraud/nonfraud로 구분했을 때 nonfraud가 반복되는 transaction 비율이 전체랑 일치하는데 fraud를 봤을 때는 c차이가 있다.

카드당 반복되는 거래의 비율로 보면 fraud만 보면 outlier들이 생겨서 반복거래의 패턴이 깨진다.

'반복거래'

어떻게 데이터로 표현을 할까? 그냥 Boolean으로? -> test할 때는 없다. Boolean으로 하기 힘들 것

그럼 Fraud일 때 반복거래에 무조건 Boolean으로 1을 주자!

아니면 반복거래에 column을 주면 되지 않나?

전체 거래에서 반복거래의 비율을 주자! 이게 더 좋을 것이다.

한 카드 identity마다 주는 것이 어떻냐?

**# 거래비중 Top3**

시간대가 같다. 아마 미국, 멕시코, 캐나다 일 것이다.

이들 사이에서 fraud의 관계를 보자.

**# Amt**

소수점 15자리까지 내려가는 것이 보인다.

**# 희영이가 할 말이 많다요**

V101~V339까지 봤다.

Null과 nonnull로 유의미한 것은 단 하나의 column으로 줄일 수 있다.

corr는 0.13정도

null를 완전 제거하고 nonnull로 corr로만 계산했을 때 0.36이 나왔다.

즉, null이 있는 상태에서 적용하기는 힘들다.

여러 가지를 생각해봤다.

Imputation을 KNN, DL로 할까 생각해보다가

KNN으로 하면 안되고 median이 mean보다 imputation하는 게 더 낮다.

근데 이러면 상관관계가 0.09 정도 떨어진다.

PCA로 차원축소를 했다. 201개의 column이 dimension이 2개로 준다.

PCA로 만든 column 2개로 Random forest를 돌려봐서 설명력이 있으면 살리고 없으면 corr일일이 뽑은 애들 중 높은 애들만 써야할 것 같다.

PCA를 효과적으로 쓰려면 아예 희영이가 V-column을 다 해야 한다.

**# Product CD**

corr이 거의 안 나온다.

희영이가 해보니까 나오는데?

1개씩 하면 나오던데? 2개는 나오고 2개는 안 나올껄?

다시 소영이가 확인해봐야 합니다.

one-hot encoding으로

**# 발표**

Card3가 국가인거 밝히고

높은 거 몇 개만 공개

column은 공개할 수는 없지만 이런 column도 있을 것이다.

데이터를 잘 써서 실제 시간을 알아보니

나라 3개 시간 똑같은 거. 미국, 멕시코, 캐나다일 것이다.

있어bility

시각화를 잘해야 한다.

F1 score

우리가 내일까지 column을 전처리 과정을 다 거쳐서 현수누나가 돌릴 수 있게 보내줘야 한다.

Column을 잘 선택해서 줘야할 것 같다.

각자 맡은 column export해서 보내줘야 된다아아아아!

dataframe 다시 다 export해서 현수누나한테 보내줘야 하겠네.

- column을 왜 바꿨는지 설명을 해줘야 한다.

희영이가 V-column을 다 하면 현수누가가 C, D-column을 맡아줄 수 있는가?

**# One-hot encoding을 위한 노가다 Code**

이미 있었던 code였던 걸로 판별

**# 중간발표 전까지 어떻게 진행을 할 것인가**

언니가 수요일 저녁에 올리려면 오는대로 돌려보면 된다.

수요일 저녁 6시전까지 보내주는 걸 마지노선으로 하자.

**# 희영**

PCA 축소한 것, 임의로 고른 행렬 있는 것, 여러 개를 보내줄게!!

**# Device type**

진우가 할 것이다. 그걸 출시일자는 절대 못한다.

**# 가장 중요한 것**

같은 금액으로 해서 움직이는 것은 어떻게 할 것인가

# 새로 만들어지는 column 개수는 대략 어느 정도 될까?

새로 만드는 것만?

Amt Fraud 비율 1개

국가별 지역별로 사기위험은 비율값으로 대체

시간도 country로 나눠서 비율값으로 대체. Fraud의 비율인데 test에는 fraud가 어떻게 모른다.

transaction

#

시간에 따른 사기비율

실제 데이터에서는 이 column을 어떻게 분석을 할 것인가.

Fraud비율자체가 Fraud를 알 수 있을 때만 나올 수 있는 것 아닌가?

지금 우리 가설

e.g. 미국에서 13시에 5%이고 14시에 10%의 fraud비율이 나온다.

그럼

# 각자 할 일들

내가 할 일. 시각화!

희영: 왜 imputation을 이렇게 하게 했는지 표를

현수: 누나는 모델을 돌리고

현우: 있어bility를 올리기 위해서 어떻게 시각화하면 좋을지 고민하기

해당 시간에 따른 사기 비율 / 전체 사기 비율