**PPT 1. (다른 표지 후보는 PPT 19)**

3조 발표를 시작하겠습니다.

**PPT 2**

발표 목차는 크게 데이터가 어떤 특징을 가지고 있는지, 데이터를 어떻게 전처리했는지, 사용 모델은 무엇을 했고 이를 통한 성능은 어느 정도가 나왔는지 순으로 진행하겠습니다.

**PPT 3**

우선 전체적인 데이터의 특징을 말씀드리겠습니다.

**PPT 4**

이번 데이터는 크게 identity와 transaction Data가 있고

이를 통해서 Test data의 사기거래 판단이 이번 프로젝트의 목적입니다..

identity는 row가 14만개 정도이고, Transaction은 row가 59만개 정도입니다.

전체 거래 데이터에서 사기거래의 비중은 약 3.5%입니다.

각 column들에 대한 정보가 굉장히 부족했기 때문에 각각 column 어떤 것을 의미하는지 알아내기 힘들었습니다.

**PPT 5**

이제 저희가 각 column들 중에서 어떤 의미를 도출했으며 각 column들을 어떻게 전처리를 했는지 보여드리도록 하겠습니다.

**PPT 6**

먼저 identity column들에 대해 설명 드리겠습니다.

38개의 id 칼럼들을 일일이 다 상관관계를 보았고 imputation까지 해봤는데 corr가 거의 없었습니다.

저희 조는 id에 있는 value들의 절대적인 값들이 유의미하지 않다고 판단해서 binary encoding을 진행하여 NA=0으로 아니면 =1로 Encoding을 했습니다.

PCA를 거치고 칼럼 중에서 상관관계가 있는 것만 남겼습니다.

그리고 DeviceType, DeviceInfo는 현재 데이터로만은 유의미한 상관관계를 도출하지 못하여서 외부 데이터를 끌어와 유의미한 상관관계를 도출하기 위해 시도하고 있습니다.

`

**PPT 7**

다음은 TransactionDT와 TransactionAmt 칼럼입니다.

위 column들은 각 거래가 일어난 시간의 델타값과 한 번 거래가 일어났을 때 얼마만큼의 금액이 거래되었는지 보여줍니다.

위 column들을 바탕으로 저희는 거래 발생 시점을 도출할 수 있었습니다.

TransactionAmt에서 사기거래와 사기거래가 아닐 때의 특이점을 발견해서 이를 반영해주었습니다.

**PPT 8**

또한 저희는 TransactionAmt에서 사기거래일 때 같은 금액이 짧은 시간내에 반복적으로 거래가 되는 경향을 발견하였습니다.

이는 카드사기의 경우, 복제한 카드 혹은 훔친 카드가 잘 작동하는지 확인하기 위해서 입출금을 반복하는 경향이 있기 때문입니다.

저희 조는 이 또한 반영을 하였습니다.

**PPT 9.**

다음 어떤 종류의 상품이 거래되었는지 알려주는 ProductCD column에서는 더미변수화 시켰습니다.

**PPT 10.**

다음은 card, addr, dist, emaildomain 칼럼들입니다.

저희는 이 중에서 각 카테고리 변수와 사기거래와의 상관관계를 살펴본 뒤 이메일 column과 같이 유의미한 관계가 발견되지 않는 칼럼들은 삭제를 하였습니다.

VESTA 측의 설명에 따르면 위 card 칼럼은 card type, card category, 발행 은행, 국가 등의 정보가 담겨져 있습니다.

저희는 이 중에서 어떤 칼럼이 국가에 해당하는지 찾고자 노력했습니다.

**PPT 11.**

국가 컬럼을 찾고자 했던 이유는 사기거래와 국가 간의 상관관계가 높기 때문입니다.

논문과 발표자료에 따르면 전체 카드 사기거래의 70% 이상이 지도에서 보이는 우크라이나, 미국, 터키, 인도네시아 등 6개 국가에서 발생합니다.

이를 고려하여 저희는 각 column 중에서 어떤 column들이 국가인지 찾은 뒤, 사기거래와 일반 거래량의 추이를 비교해보았습니다.

**PPT 12.**

왼쪽 그래프는 특정 국가에서 시간대별 사기거래를 퍼센트로 나타낸 것이고

오른쪽 히스토그램은 시간대별 일반 카드거래의 추이를 나타낸 것입니다.

위처럼 사기인 거래가 일반 카드거래와는 다른 시간추이에 발생하는 경향을 포착한 뒤 이를 반영하였습니다.

**PPT 13.**

나아가 저희는 각 국가 내에서 위치, 거리에 따라서도 사기비율이 달라진다는 것을 알 수 있었습니다.

실례로 한 국가 같은 경우 거리별로 28개의 section으로 나눌 수 있었는데 이 중에서 7개 구역에서만 사기거래가 발생하고 사기 거래비율도 현저하게 높았습니다.

추가적으로 이 국가별 section을 나눠주는 column에서 NaN/NonNaN을 구분하는 바이너리 컬럼 추가했습니다.

**PPT 14.**

그 다음은 C, D, V-column 칼럼들입니다.

C는 isFraud 비율로, D는 null값이 많아서 pandas\_profiling에서 null 값이 50% 이하인 것으로 추림.

위 그래프의 선들은 각 column들 별로 isFraud 칼럼과의 상관관계를 나타내줍니다.

빨간색 선은 null값들을 mean으로 imputation을 한 것이고, 검정색 선은 median으로 imputation을 한 것입니다.

mean으로 imputation한 것보다 median으로 imputation한 게 isFraud column과의 상관관계가 더 높은 구간이 대체적으로 많은 것을 보실 수가 있습니다.

그래서 median으로 NA값들을 대체한 뒤 PCA를 거쳐서 칼럼들의 수를 줄이는 방법으로 C, D, V-column들을 처리하였습니다.

**PPT 16.**

어떤 모델을 사용했으며 어떤 성능이 나왔는지 이후에 설명하면 될 듯