TCT-기술인증테스트
데이터분석
서술형 모범답안

[2021년 #차]

사번		성명	
유의 사항	▪ 본 서술형 문제지에 제시된 시나리오와 문제를 읽고, 별도 배포되는 답안지에 응시자의 답안을 작성해 주시기 바랍니다. ▪ 배포된 문제지는 시험 종료 후 답안지와 함께 감독관에게 제출해 주시기 바랍니다. ▪ 공정한 평가를 위하여 답안 작성 시 동료를 도와 주는 행위, 보여주는 행위를 금지하고 있으며, 평가에 대한 부정행위 적발 시, 응시한 평가는 0점 처리됩니다.		

LG CNS

[서술형 1번] 금융거래 이상감지(24점)

A 카드사는 신용카드 부정 사용을 적발하기 위해 머신러닝 기법을 활용하고자 한다. 모델 개발을 위해 고객 정보 데이터와 거래이력 데이터를 이용할 수 있다. 2020년 1월1일부터 2021년 6월30일까지 총 1년 6개월 간의 데이터를 분석에 이용할 수 있다.

[데이터 개요]

- 고객정보 데이터

column	column 설명
Customer ID	고객별 고유번호
Gender	성별(여성, 남성)
Age	연령대(20대, 30대...)
Marital status	결혼여부(기혼, 미혼)
Children	자녀 수(0,1,...)
Region	거주지역(구 단위)
Job	직업(회사원, 전문직...)
Salary	연소득(단위:만원)

- 거래이력 데이터

column	column 설명
Customer ID	고객별 고유번호
Datetime	거래일시(YY-MM-DD hh:mm:ss)
Business Category	가맹점 업종
Location	가맹점 위치
Amount	결제 금액(단위:만원)
isFraud	부정사용여부

[문제 1-1] 분석 마트 설계(5점)

고객의 카드 사용 특성과 부정사용 여부의 연관성을 통해 **부정사용 건을 적발**할 수 있는 모델을 만들려고 한다. 이를 위한 데이터 마트를 설계하시오. (모델에 들어갈 독립변수, 종속변수를 정의하고, 행의 기준(key)을 정의할 것)

[문제 1-2] 모델설계(7점)

위에서 설계한 데이터 마트를 이용하여 부정사용 건을 적발하기 위한 방법론을 정의하시오.
(즉, train/test 분리방법, 사용할 알고리즘(모델)과 선정이유, 모델성능 검증방안 등을 쓰시오.)

[문제 1-3] 원인도출(5점)

위에서 설계한 방법론 대로 신규 데이터에 대해 검증을 수행하던 중, 부정사용 의심 건이 도출되었다. 이 건에 대해 어떤 점이 부정사용으로 판단하는데 기여하였는지 확인할 수 있는 방법론을 정의하시오.

[문제 1-4] 모델설계(7점)

고객별로 **정상 사용 패턴과 차이가 큰 경우에 적발**할 수 있는 모델을 만들려고 한다. 이를 위한 데이터 마트 와 모델을 설계하시오.

[서술형 1번] 금융거래이상감지

[문제 1-1] 학습용 데이터 구성

<모범 답안>
[Key]
각 개별 카드 사용건 마다 부정사용 여부를 판정해야 하므로 행의 기준은 사용건이 되어야 하며 key는 Datetime과 CustomerID이다.
[독립변수와 종속변수]
- 파생변수: costomerID별 결제금액의 건별평균/표준편차를 산출하여 결제금액을 표준화
  최근 1,3,6,12개월간 결제금액평균/표준편차, 결제건수평균/표준편차등
- 독립변수: 거래이력데이터와 고객정보데이터를 CustomerID를 기준으로 병합하고,
  개별 사용건의 정보를 담고 있는 거래이력데이터의 business category, location,
  amount와 사용자의 정보를 담은 고객정보데이터의 gender, age, marital status,
  children, region, job, salary, 위에서 정의한 파생변수들을 독립변수로 사용함
- 종속변수: isFraud

[문제 1-2] 모델링 방법론 정의

<모범 답안>
- train/test분리 : train/test셋은 시간 기준으로 분리를 해야함. 또한 train에서 표준화된 파생변수를 정의하였으므로, test 셋의 파생변수는 train 셋의 평균과 표준편차를 이용하여 계산되어야 함.
- 모델 : 종속변수가 이항분포를 따르는 isFraud이므로 logistic regression, Randomforest, Xgboost, Neural Network(DNN), SVM, KNN등 classification models을 사용할 수 있다.
- 모델성능검증방안 : 모델의 검증 방안은 여러 개 모델들에 대해 동일한 test셋에 대한 평가지표로 모델성능을 비교하고, 이때 모델이 실제 부정사용건을 놓치면 안되기 때문에, recall을 100%(혹은 95,98%등 매우 높은 수준)로 고정하고 precision을 최대화 할 수 있는 모델을 만들어야 한다. 따라서 recall을 동일한 수준으로 맞춘후 precision을 비교하여 가장 높은 precision을 가진 모델을 선정한다. 또한, 시간에 따른 모델이므로 학습 기간 및 테스트 기간을 moving하며 여러 번 검증하여 robust한 모델을 만드는 것이 필요함.

[문제 1-3] 원인도출

<모범 답안>
신규데이터 개별 사용건별로 SHAP explanation force plot을 산출하면 시각적으로 어떤 인자가 +혹은 -로 기여하여 이런 예측값이 도출되었는지를 판단할 수 있다.

[문제 1-4] 모델설계

<모범 답안>
1-1번에서 제시한 데이터마트에서 정상데이터(isFraud=0)인 행만 선별하여 이상탐지모델을 만들고, 정상데이터에서 많이 벗어난 유형의 데이터가 신규로 유입될 경우 이를 부정사용 의심 건으로 알람을 띄우도록 한다. 이상탐지 모델은 PCA-hoteling's T2/SPE, Autoencoder input, output layer의 차이, one-class SVM, isolation Forest등 정상데이터만으로 정상패턴을 추출하고 이와 벗어난 데이터를 선별하는 모델들을 사용할 수 있다.

[서술형 2번] 고객 등급 Classification 문제 해결 (24점)

L사는 각 고객들의 가입기간, 서비스 이용 규모 등을 통해 고객을 총 3등급으로 나누어서 관리하고 있다. L사에서는 최근 데이터를 활용한 고객을 선제적으로 관리하는 trend에 따라 3개월 후 고객 등급을 예측하여 등급이 우수할 것으로 예상되는 고객들을 찾고자 한다.

이를 위해 L회사 내 분석가 조직은 **각 고객별 과거 3개월간** L사에서 제공하는 모든 서비스 이용 이력을 활용하여 이번 달 고객 등급을 정답으로 하여 분석 모델링을 구축하고자 한다.

[문제 2-1] 분석 설계 (4점)

위 데이터 상황과 목적 하에서 L사 분석가 조직에서는 다음과 같이 데이터셋을 만들고, 모델링 구축을 진행하였다.

- 2020년 12월을 기준으로 과거 3개월간 고객들의 서비스 이용 이력을 활용하여 분석함
- 최대한 많은 파생변수를 만들어서, 활용할 수 있는 모든 변수들과 함께 모두 모델링의 feature로 반영하여 모델링을 구축함

분석하고자 하는 의도를 참고하여, 위와 같이 데이터셋을 구축하였을 때의 문제점을 위의 각 불릿별로 1가지 이상 서술하고, 아래 문항들을 해결하기 위한 데이터셋 구성 방안을 서술하시오.

[문제 2-2] 추정을 위한 모델링 및 검증 - 1 (6점)

L사 분석가 조직에서는 먼저 로지스틱 회귀분석을 이용하여 모델링을 하고자 한다. 회귀분석식을 제시하고, 선형성 가정에 대한 확인하는 방법을 서술하여야 하며, 2-1에서 제기한 문제점을 고려하여 로지스틱 회귀분석을 활용하여 모델링하고, 등급 여부를 판단하는 방안을 서술하시오. (각 고객 등급별 여부를 예측하는 one-to-all로 3개의 모델링을 구축한다는 가정 하에 특정 등급 여부를 예측하는 방안으로 서술. 변수는 임의로 $x_1, x_2, \cdots$로 표현하여도 무방함)

[문제 2-3] 추정을 위한 모델링 및 검증 – 2 (8점)

L사 분석가 조직에서는 동일한 feature를 사용한다고 할 때, 로지스틱 회귀분석에서는 반영하기 힘든 효과가 있어 tree기반의 모델링을 하기로 하였다. 이 효과가 무엇인지 설명하고, tree기반 모델링 중 하나를 선정하고, 모델링 구축과정과 모델 검증하는 방안을 상세하게 서술하시오.

[문제 2-4] 모델 평가 (6점)

모델링 결과, 다음과 같은 Confusion Matrix를 도출하였다.

예측 \ 실제	등급 1	등급 2	등급 3
등급 1	a	b	c
등급 2	d	e	f
등급 3	g	h	i

당신이 L사 분석가 조직 내 분석가라면, Multi-class라는 점을 고려하여, Confusion Matrix에서 도출할 수 있는 평가 지표 중 Accuracy를 제외하고, 지표 3가지 이상 정의하고, 실제 그 지표 값을 산출하는 과정을 서술하시오.

[서술형 2번] **고객 등급 Classification 문제 해결**

[문제 2-1] 분석 설계 (4점)

<모범 답안>

- 일률적으로 3개월로 기간을 설정하여 데이터셋을 구성할 경우, 각 고객별로 반영될 수 있는 기간이 달라질 수 있다.
- 최대한 많은 파생변수를 만드는 것 자체는 문제가 없지만, 모두 모델링에 반영할 경우, Overfitting의 문제가 발생할 수 있다.
- 각 고객별로 가장 최근의 사용이력이 있는 일자를 기준으로 각각 과거 3개월까지의 사용이력을 요약할 수 있도록 데이터셋을 구성할 것이며, 최대한 많은 파생변수로 구성하되 모델링 과정에서 변수 선택 등의 과정을 진행하고자 한다.

[문제 2-2] 추정을 위한 모델링 및 검증 - 1 (6점)

<모범 답안>

- 회귀분석식은 다음과 같이 정리할 수 있다.

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$$

- 위 식에서 Least Square Estimate를 만족하는 베타값들을 찾도록 하되, 모든 변수들을 다 사용하면, 예측력에 문제가 있을 수 있으므로, stepwise 등의 방법을 통해 주요 변수들을 선택하고 회귀식을 만들어야 한다.
- 선형성의 가정은 회귀모형식에서 x값들과 y가 서로 선형 관계임을 가정하는 것으로 위 식에 따라 확률값의 logit 함수 값과 x값들이 서로 선형 관계인지를 확인하는 것으로 plotting을 통해 확인할 수 있다.

- 최종적으로 고객 등급 여부는 고객 등급에 포함되는 확률 p값을 계산하고, 0.5이상이면 고객 등급에 포함되는 고객으로 판단한다.
  (p값에 따른 recall / precision 등에 따라 기준값이 달라질 수는 있지만, 일반적으로 0.5를 기준으로 판단하면 크게 무리는 없다.)

[서술형 2번] **고객 등급 Classification 문제 해결**

[문제 2-3] 추정을 위한 모델링 및 검증 – 2 (8점)

<모범 답안>

- 일반 선형회귀분석에서는 교호 작용이 반영되기 어렵다는 단점이 있어 Tree기반의 알고리즘을 활용하면 좀더 나은 결과를 기대할 수 있다.
- Tree 기반 알고리즘 중 RandomForest를 사용하려고 하며, RandomForest를 통해 모델링을 구축함에 있어, tree 깊이와 한번에 변수 몇 개를 random하게 추출할 것인지를 주요하게 hyper parameter 튜닝을 하고자 한다.
- 모델링 구축 후 train set에서의 loss값과 validation set에서의 loss 값을 비교하면서 오버/언더피팅을 진단하여 최종 하이퍼파라미터를 정하여 최종 모델링을 완성한다.

[문제 2-4] 모델 평가 (6점)

<모범 답안>

- Micro-precision
- Micro-recall
- Micro-F1
- Macro-precision
- Macro-recall
- Macro-F1 등

수고하셨습니다.