Week06 발표

Dacon 이상신용거래탐지 대회 4위 코드 리뷰

이경선

4위 코드 공유의 방법

안녕하세요.
대회에 참여하신 분들 모두 고생 많으셨습니다.^^

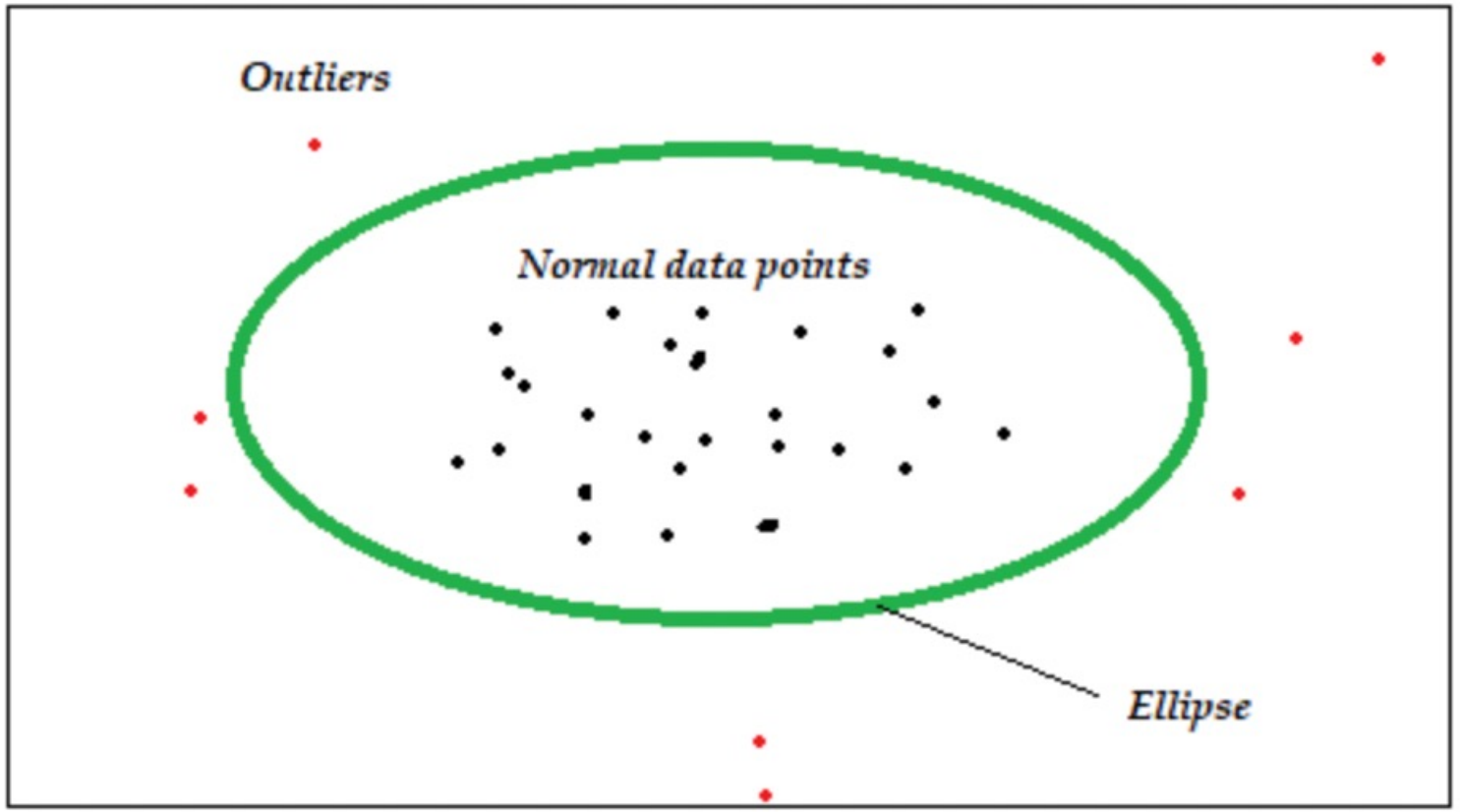
저희 팀은 코드 공유 게시판에서 참조한 EllipticEnvelope으로 첫번째 예측값을 얻고,
EllipticEnvelope으로 trainset에 임의 label을 주어 LGBM으로 모델을 최적화하여 두번째 예측값을 얻은 뒤
Ensemble하는 방식으로 간단하게 결과를 냈습니다.

특이사항으로 AorB:True 방식으로 Ensemble을 한 이유는
신용카드 사기 거래 탐지는, 한번이라도 fraud로 예측된 example은 fraud로 결정하는 보수적인 의사 결정이 필

감사합니다.

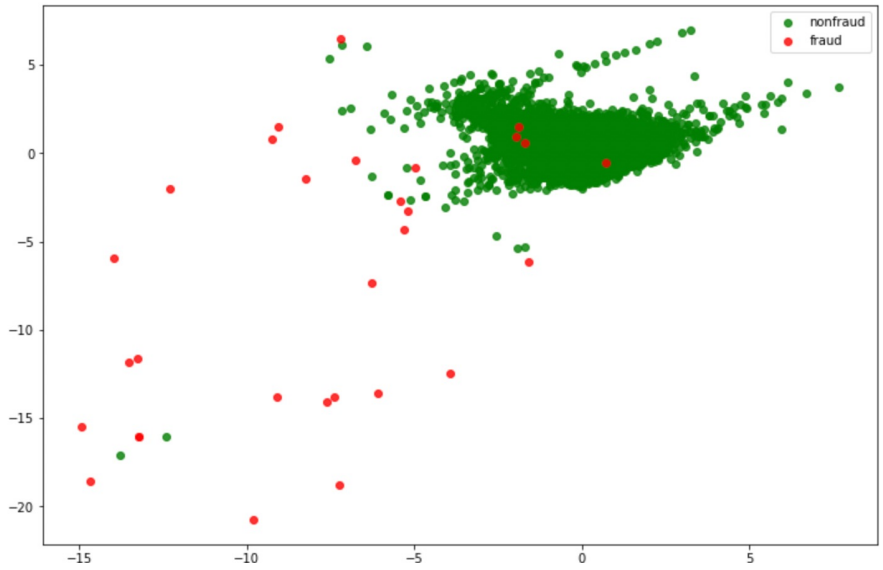
EllipiticEnvelop: label 만들기 〉 LGBM: 모델 최적화 〉 AroB:True: 앙상블

EllipiticEnvelop

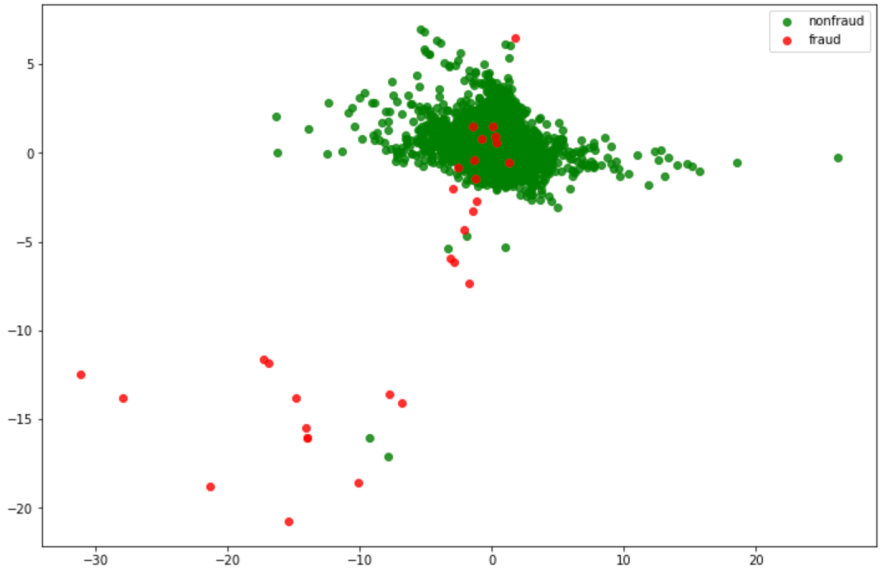


타원 밖의 점들은 모두 Outlier로 판단하는 모델.
데이터가 가우스 분포를 가질 때 가장 잘 작동한다.

Dacon data visualization



Column "V14","V17" 시각화



Column "V7", "V17" 시각화

EllipiticEnvelop

```
model = EllipticEnvelope(support_fraction = 0.994, contamination = fraud_ratio, random_state = 42)
model.fit(trainset)
```

support_fraction : *float, default=None*

The proportion of points to be included in the support of the raw MCD estimate. If None, the minimum value of support_fraction will be used within the algorithm: `[n_sample + n_features + 1] / 2`. Range is (0, 1).

contamination : *float, default=0.1*

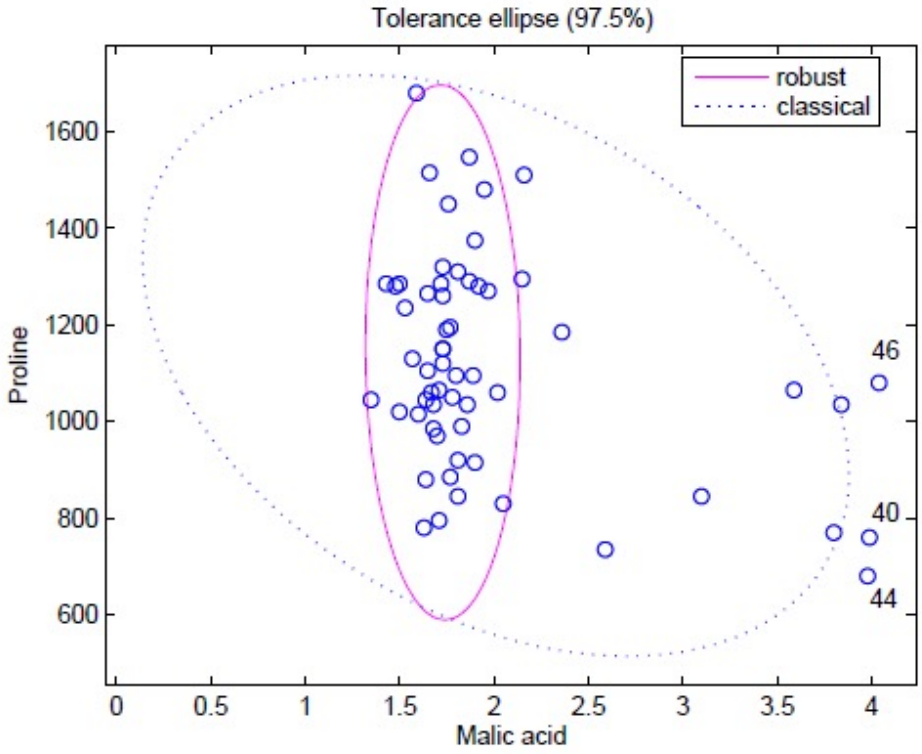
The amount of contamination of the data set, i.e. the proportion of outliers in the data set. Range is (0, 0.5].

random_state : *int, RandomState instance or None, default=None*

Determines the pseudo random number generator for shuffling the data. Pass an int for reproducible results across multiple function calls. See Glossary.

Support_fraction: raw MCD estimator의 비율

MCD: Minimum covariance determinant



데이터에서 h개의 샘플을 뽑고, 공분산이 가장 작은 데이터를 선정한다.
선정된 데이터의 평균, 분산의 Mahalanobis 거리를 산정한다.
이상치 탐색에 적절한 특징이 있다.

Mahalanobis 거리

$$d(u, v) = \sqrt{(u - v)\Sigma^{-1}(u - v)^T}$$

covariance matrix의 inverse matrix를 곱하여 거리를 재는 방식이다.
이를 통해 변수들간의 correlation등 분포를 고려하여 거리를 잴 수 있다.

$$d_{\Sigma}^2(X_i, \mu) \sim \chi_p^2$$

데이터가 다변량 정규분포를 따른다는 가정 하에서
변수가 p개인 마하 거리의 제곱은 카이제곱 분포를 따른다.