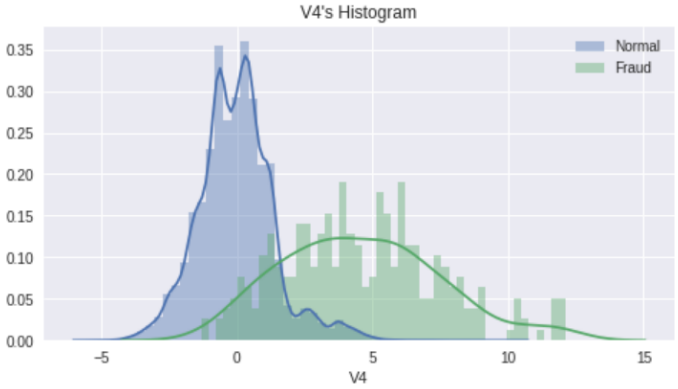
1. Problem Definition

총 49,200 개의 신용 카드 거래 내역 중 1%인 492 개의 거래는 비정상 거래이고, 나머지 48,708 개의 거래는 정상 거래이다. 그 중 비정상 거래를 찾아내는 기계학습 모델을 만드는 것이 목표이다.

2. EDA




```
The total counts of each class are  28908 for class 0 and  292  for class 1
```

먼저 EDA 를 통해 데이터가 어떤 분포를 띄고 있고, 어떤 특성을 보이는지 '감'을 잡는 것이 중요하다. Training set 을 살펴보면 29,199 개의 거래 내역 중 28,908 개가 정상 거래, 292 개가 비정상 거래로서 클래스 불균형이 매우 심했다. 클래스 불균형이 심한 분류 문제를 풀 때 어떤 방법을 사용해야 하는지 고민이 필요함을 알 수 있다. 다행히도 결측치는 없어서 imputation 은 필요하지 않았다.

우리가 알고 싶은 것은 비정상 거래이므로 비정상 거래를 class 1, 정상 거래를 class 0 으로 나눈 후 함께 plot 을 해보니 여러 feature 에 대해서 class 간 차이가 심한 것이 있었다. 해당 feature 들이 class 를 분류하는데 key point 가 될 수 있음을 예상해볼 수 있다.

3. Proprocessing

class 0 과 1 간 분포의 차이가 거의 같은 feature 는 분류에 도움이 되지 않을 것이라 판단하여 'V28', 'V27', 'V26', 'V25', 'V24', 'V23', 'V22', 'V20', 'V15', 'V13', 'V8'를 drop 했다. 또한 Amount 를 제외한 모든 feature 가 normalization 되어 있었으므로, scale effect 를 제거하기 위해 amount 도 normalize 하였다.

4. Resampling & Model Learning

클래스 불균형을 해소하기 위한 여러 방법은 resampling을 사용하였다. Class 1의 수가 지나치게 적어 class 0에 비해 학습이 덜 되는 것을 방지하기 위해 upsampling 방법인 SMOTE를 채택했다.

Model은 Random Forest를 사용하기로 했다. Random Forest는 Ensemble기법을 사용한 모델로서 주어진 데이터로부터 여러 개의 모델을 학습한 다음, 예측 시 여러 모델의 예측 결과들을 종합해 사용하여 정확도를 높이는 기법이다. Random Forest는 두 가지 방법을 사용해 다양한 의사 결정 나무를 만드는데, 첫 번째는 의사 결정 나무를 만들 때 데이터의 일부를 복원 추출로 꺼내고 해당 데이터에 대해서만 의사 결정 나무를 만드는 방식이다. 즉, 각 의사 결정 나무는 데이터의 일부만을 사용해 만들어진다. 두 번째는 노드 내 데이터를 자식 노드로 나누는 기준을 정할 때 전체 변수가 아니라 일부 변수만 대상으로 하여 가지를 나눌 기준을 찾는 방법이다.

새로운 데이터에 대한 예측을 수행할 때는 여러 개의 의사 결정 나무가 내놓은 예측 결과를 투표voting 방식으로 합한다. 예를 들어, 총 5개의 의사 결정 나무 중 Y를 예측한 나무가 3개, N을 예측한 나무가 2개면 Y를 최종 결과로 결정하는 방식이다. Random Forest는 고차원 데이터에서 분류 문제를 푸는데 탁월한 성능을 보이는 것으로 알려져 있고, 의사 결정 나무 하나가 아니라 여러 개를 사용해 over-fitting 피한다.

5. Hyper Parameter Setting with Validation Set

```
Results with 100 estimators and None max_depth

Validation
recall:  0.81
precision:  0.941860465116
f1 measure:  0.870967741935
```

validation set을 통해 다양한 hyper parameter를 주며 모델을 평가해보았다. 결과적으로 왼쪽에서 보이는 바와 같이, 나무의 개수인 n_estimator는 100개, max_depth는 주지 않았을 때 가장 높은 f1 score를 보였다.

6.아쉬운 점과 Future Study

조성준 데이터마이닝 담당교수님의 조언에 따라 "Gaussian Mixture Model을 이용한 Anomaly Detection"으로 접근해보았다. 하지만 recall에 비해 precision이 너무 낮아 형편없는 f1 score를 보였다. 학습이 제대로 되지 않은 것으로 추측된다. 시간이 부족하여 이번에는 더 시도하지 못했지만 추후 해결해보고 싶은 문제다.

random forest는 훈련데이터의 변화에 따라 선택되는 변수군이 크게 달라지는 단점이 있으며, decision tree의 계층적 구조로 인해 불필요한 변수들도 함께 선택될 수 있다. 이는 숲을 나무의 수가 많아질수록 악화된다. 또한, Regularized RF와 Guided RRF 분류모델은 최소한의 변수군으로 모델이 구성되기 때문에 GRRF와 RRF로 중요변수를 선택하고 Random Forest로 분류 모델을 구축하는 방법이 추천된다는 것을 알게 되었다.
이번 문제에서는 변수선택을 내가 임의적으로 했지만 추후 위의 두 방법을 도입하여 성능을 더욱 향상 시켜보고 싶다.