**Machine Learning Assignment**

진 수 정

2017311974

통계학과

1. **Problem Definition**해결하고자 하는 문제에 대한 정의

주어진 문제는 각각의 비디오에 등장하는 사람의 음성, 시각 정보를 통해 해당하는 사람이 우울증을 가지고 있는지, 가지고 있지 않은지를 예측하는 Binary Classification 문제이다.

1. **Data**
   1. **Data Description**데이터에 대한 자세한 명세(형태, 피처, 클래스의 개수 등)

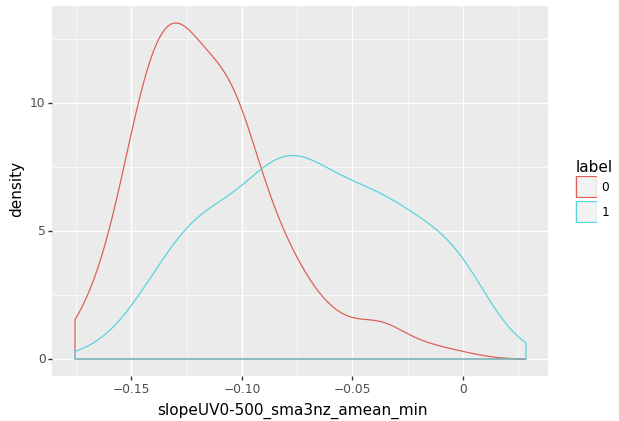
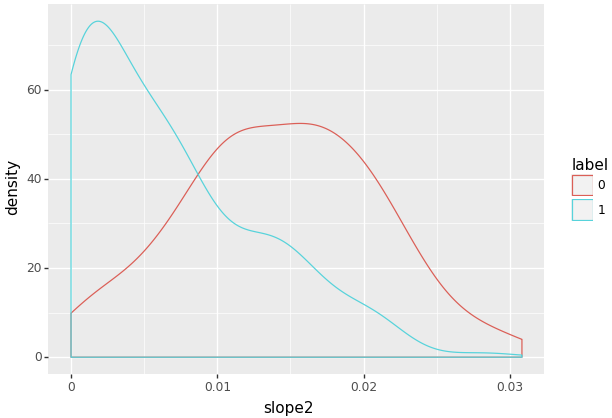
train.csv: 405 \* 384 사이즈의 데이터프레임 형태이다. 피처들은 음성, 시각 데이터의 평균, 분산, 최소값, 최대값 등의 수치형 데이터로 구성되어 있다. 우울증 여부를 나타내는 타겟은 2개의 클래스를 가지는 범주형 데이터이다. 이 때 타겟의 각 클래스의 비율이 크게 차이가 나지 않으므로 해당 데이터는 balanced data이다.

X\_test.csv: 100 \* 383 사이즈의 데이터프레임 형태이다. 피처의 개수와 형태는 train 데이터셋과 동일하며, test 데이터셋에는 타겟 변수가 누락되어 있다.

Train 데이터셋과 test 데이터셋에서 결측치는 없다. 다만 변수의 개수가 383개로 많기 때문에, 추후에 변수 선택이나 PCA를 통한 차원축소 등의 작업이 필요해 보인다.

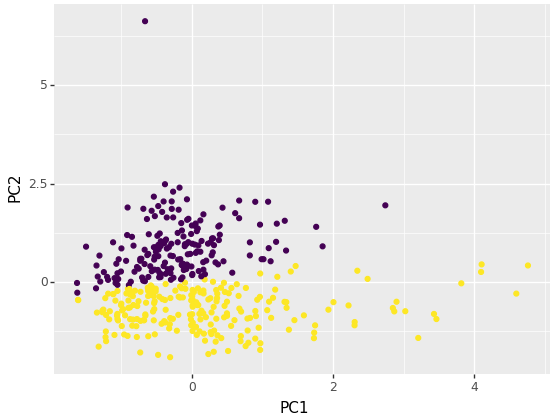
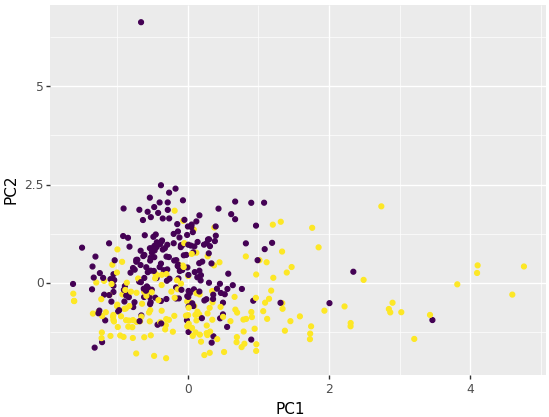
* 1. **Data Pre-processing**텍스트 전처리 등 자세한 과정 설명
* Uniqueness: 피처의 유니크 값 개수를 카운트하여, 하나의 값으로만 이루어진 7개의 피처를 확인하였다. 해당 피처들은 아무런 정보를 포함하고 있지 않는 피처이므로 제거하였다.
* Correlation coefficient: 피처들의 상관계수가 높으면 다중공선성 문제가 발생하여 예측을 잘 하지 못할 수 있다. 해당 데이터에서 모든 피처들이 수치형 데이터이므로 피처들 간의 피어슨 상관계수를 계산하였다. 상관계수 절댓값이 0.9보다 높은 변수쌍에 대하여, 라벨과의 상관계수가 더 낮은 변수를 제거하였다. 피처의 개수가 377개에서 192개로 줄었으며, 이를 통해 예측 성능 향상과 모델링 시간 감축을 기대해볼 수 있다.
* PCA: Train set과 Test set의 피처들에 대하여 정규화를 시켜준 다음 PCA를 진행하여 5개의 새로운 변수를 얻을 수 있었다. 5개의 주성분으로 데이터의 42% 정도가 설명이 되기 때문에 데이터를 5개의 주성분으로 축소하기에는 무리가 있지만, 해당 변수들을 통해 기존의 변수로는 설명할 수 없었던 데이터의 일부가 설명될 수도 있다고 판단하였다. 따라서 5개의 주성분을 데이터셋에 새롭게 추가하였다.

1. **Analysis**모델에 활용될 피처에 대한 분석(Word Count 등)

* Visualization & Feature Engineering

라벨과의 상관관계가 높았던 피처 'slopeUV0-500\_sma3nz\_amean\_min’에 대하여 라벨별로 'slopeUV0-500\_sma3nz\_amean\_min’의 분포를 시각화 해보았다 (왼쪽). 시각화 결과, 라벨별로 피처의 분포가 잘 구분되므로 해당 피처는 라벨 분류에 효과적인 피처라고 볼 수 있다. 또한 해당 피처를 제곱한 값에 대해서도 시각화를 진행해본 결과 (오른쪽), 라벨 분류에 도움이 될 수 있을 거라 판단해서 데이터셋에 추가해주었다.

* Clustering & Feature Engineering



* PCA를 통해 추가된 피처 PC1과 PC2를 이용하여 K-means 클러스터링을 실행해 보았다. 먼저 PC1과 PC2 축에 대해 데이터의 라벨 값을 시각화 해본 결과, 데이터가 일부 overlapped 된 부분이 있음을 확인하였다 (왼쪽). PC1과 PC2 축에 대해 클러스터링 결과를 시각화 해본 결과 (오른쪽), overlapped 된 부분에서는 클러스터링 결과와 실제 라벨이 일치하지 않았지만, 클러스터링 결과가 라벨과 대략 72% 정도 일치하는 것을 확인할 수 있었다. 따라서 클러스터링 결과도 라벨 분류에 도움이 될 수 있을 거라 판단하였고 하나의 피처로 추가해주었다.

1. **Model**
   1. **Classifier**사용된 모델 설명(알고리즘, 동작 방식 등)

- Logistic Regression

선형 회귀 방식을 기반으로 하되 시그모이드 함수를 이용해 분류를 수행하는 모델이다. 세부적으로 보면, 시그모이드 함수의 최적선을 찾고 시그모이드 반환값을 확률로 간주해, 확률에 따라 분류를 결정하는 방법이다. 로지스틱 회귀는 가볍고 빠른 모델이지만, 복잡한 문제의 경우에는 다른 모델의 성능이 더 좋을 수 있다.

- SVC

training observation과 hyperplane 간의 최소거리인 margin을 최대로 하는 hyperplane을 찾는 문제이다. 보통 예측 정확도가 높으며 soft margin을 사용하여 non-separable한 케이스도 해결할 수 있지만, 연산이 많기 때문에 학습 속도가 느리다.

* Random Forest

Bootstrap으로 여러 개의 subset data를 샘플링한 후에, 개별적인 트리가 각각의 데이터셋에 대해 학습을 수행한 뒤 보팅을 통해 최종적인 예측을 하는 알고리즘을 말한다. 전반적으로 다른 머신러닝 알고리즘보다 뛰어난 성능을 보이는 경우가 많다.

* 1. **Design Consideration**디자인 고려사항

데이터셋에 피처들의 개수가 여전히 많은 것을 확인할 수 있는데, SVC나 Random Forest와는 달리 특히 로지스틱 모델의 경우 피처들이 highly correlated 되어 있을 때 예측이 제대로 이루어지지 않을 수 있다. 또한 SVC나 Random Forest의 경우에도 Computing time을 고려해봤을 때, 중요한 피처들을 선별하여 모델링하는 것이 적절하다고 판단하였다. 피처들을 선별하기 위해 먼저 라벨과의 상관계수의 절댓값이 대략 0.3 이상인 피처들을 후보로 설정하였고, 피처 후보들에 대해서 몇 가지 피처들의 조합을 naive하게 비교하여 최종적으로 사용할 피처들을 선별하였다. 실제로 피처들을 선별하지 않았을 때와 비교하여, 피처들을 선별하였을 때 모델의 예측 성능이 확연히 높아지는 것을 확인할 수 있었다.

1. **Experiments**
   1. **Settings**실험 설정(Train Test 비율, 파라미터 선정 방식, 학습 프로세스 등)

- Split Train, Validation set

Train 데이터를 Train set과 Validation set으로 분할할 때, 라벨 클래스의 비율을 비슷하게 유지하기 위해서 층화추출을 통해 데이터를 분할하였다. 또한 Train set과 Validation set의 비율은 7:3으로 정하였다.

- 파라미터 선정 방식

모델의 파라미터 각각에 대해 넓은 범위에서 그리드 서치를 하여 최적 파라미터의 대략적인 범위를 파악한 후에, 그 근방에 대하여 보다 촘촘하게 설정한 그리드에서 다시 한번 그리드 서치를 진행하여 최적 파라미터 위치를 파악하였다.

- 학습 프로세스

각각의 3가지 모델(Logistic Regression, SVM, Random Forest)에 대해 그리드 서치를 통해 얻은 최적 파라미터에서의 정확도를 비교하여 가장 높은 정확도를 보인 모델을 최종 모델을 결정하였다. 그 후, 최종 모델에 대해 세분화된 범위에서 그리드 서치를 추가로 진행하여 최종 파라미터를 구한 후, 전체 Train set으로 해당 모델을 학습시켰다.

* 1. **Performance Metrics**평가 지표 설명

Accuracy: 전체 경우에 대하여 옳게 예측한 경우의 비율을 나타내는 값으로, 직관적이기 때문에 보편적으로 사용되는 지표이다. 추가적으로 클래스 불균형이 존재할 경우 large class를 잘 예측하는 방향으로 학습이 되며 small class는 잘 예측하지 못하는 경우가 생기기 때문에, F1-score 등 다른 평가 지표가 사용되기도 한다. 주어진 문제의 경우 클래스 불균형은 없었으므로 Accuracy는 평가지표로써 적절하다.

* 1. **Results**결과 및 해석

|  |  |
| --- | --- |
| Best Model | (Validation) Accuracy |
| Logistic Regression | 0.811475 |
| SVM | 0.688525 |
| Random Forest | 0.737705 |

가장 높은 정확도를 보인 로지스틱 모델을 최종 모델로 선정하였다. 세분화된 그리드 서치 결과, 최적의 파라미터 조합은 learning\_rate 0.014, iteration 100000으로 구할 수 있었다. 전체 Train 데이터셋을 최적의 파라미터 조합에서 학습시킨 로지스틱 모델은 Test 데이터셋에 대해 83.8384의 정확도를 보여주었다 (리더보드 결과).

1. **Discussion and Limitation**활용방안 및 한계점

- 활용방안: 우울증은 슬픔, 외로움 등의 감정에 빠지게 하며, 업무에 집중을 하지 못하도록 집중을 저하시키는 등 부정적인 영향을 미칠 수 있다. 따라서 개인이 우울증을 앓고 있는지 판별하는 작업은 중요한데, 위 과제에서 구현한 Depression Detection 알고리즘을 적용해볼 수 있을 것이다. 비디오를 시청하고 있는 사용자의 표정이나 음성 데이터를 웹캠이나 오디오를 통해 실시간으로 제공받고, Depression Detection 알고리즘을 적용해 봄으로써 우울증을 판별하고, 판별 결과와 여러 통계결과를 사용자에게 제공해주는 시스템으로 구현해볼 수 있을 것이다.

- 한계점: 모델 구현의 한계로 인해 보다 좋은 예측 성능을 낼 수 있는 다양한 모델들을 시도해볼 수 없었다. 또한 모델링 시간이 오래 걸리다 보니 시간적인 한계로 인해 파라미터 튜닝을 충분히 진행하지 못했다. 추가적으로 피처의 개수가 너무 많아 개별적인 피처에 대해 탐색하기에는 무리가 있었다.

1. **References**

권철민, 『파이썬 머신러닝 완벽 가이드』, 위키북스, 2019.