○ 개요

1. 요약

분석의 첫 단계로 분석에 영향을 주는 결측치의 처리에 대해, 너무 많은 결측치를 대체할 시 악영향을 준다고 판단하여 보수적인 관점으로 결측치가 5%가 넘어가는 변수들을 제거하였습니다. 그리고 결측치를 가지고 있는 나머지 변수들에 대해서는 결측치 대체 알고리즘인 MICE를 이용하여 결측치를 대체하였습니다.
그 후, VSRURF 방법론을 이용해 유의미한 변수들을 추출하여 해당 변수들로 분석을 위한 데이터를 재구성하는 정제 과정을 거쳤습니다. 이후, 딥러닝인 분류에는 Autoencoder 모형과 회귀에는 MLP 모형을 이용해 CASCADE 방법론을 사용해 분석 모형을 구축하였습니다.

트랙 B 데이터의 경우, 범주가 0인 데이터가 90 % 이상일 정도로 0을 제외한 나머지 범주들의 데이터의 개수가 매우 불균형한 데이터이므로, 이를 올바르게 학습하기 위해 CASCADE 방법을 이용했습니다. 먼저 “전기차 보유 수” target 데이터를 0과 1의 이진 데이터로 변환하여, y 의 범주 비율에 맞춰 train 데이터와 test 데이터를 분리했습니다. 분류와 회귀 모형 두 가지를 각 단계 별로 두고, 분류 모형 AutoencoderClassifier 의 학습을 진행하였고, 거기서 1로 분류된 데이터 포인트들에 대해 2차적으로 주최 측에서 제시한 회귀 분석의 target 변수로 변환하여 학습을 진행하였습니다. 그리고 test 데이터를 이용해 검증한 결과 좋은 예측 성능을 보이는 것을 확인했습니다.

2. 착안점

트랙 B 데이터의 경우, 범주가 0인 데이터가 90 % 이상일 정도로 0을 제외한 나머지 범주들의 데이터의 개수가 매우 불균형한 데이터입니다. 이러한 데이터를 분석할 때는 oversampling 이나 undersampling을 이용하는 것이 기초적이지만, 두 방법이 feature 의 개수가 많은 데이터에서는 치명적이라 판단했습니다. 이러한 데이터에 적합한 방법론을 고민하다가 데이터의 중요한 특성과 패턴을 자동으로 학습하는 Autoencoder 의 경우, 재구성 오류를 기반으로 이상치를 탐지하는 성질을 가지고 있으므로 소수 클래스를 더 쉽게 식별할 수 있어 이 데이터에 효과적인 모형이라는 판단을 내렸습니다. 또한, 데이터의 개수가 많은 클래스가 0에 해당하는 데이터들의 일부는 유사할 가능성이 높아 Autoencoder 가 이러한 중복성을 줄이고 데이터의 중요한 정보만을 압축된 형태로 보존할 수 있을 것이라 기대했습니다.

물론 모델이 소수 클래스의 예시가 너무 적기 때문에 모형이 충분한 정보를 얻지 못하여 소수 클래스에 대해 데이터를 잘못 재구성할 가능성이 존재한다는 단점이 있습니다. 이를 보완하기 위해 결측값에 대해 보수적인 관점을 유지하여 5% 이상을 결측치로 가진 변수들을 제거

하고, 결측치가 존재하는 나머지 각 변수들을 종속 변수로 사용하여 남은 변수들을 독립 변수로 두어 회귀 모형을 설정해 결측치를 예측하고 대체하는 MICE 알고리즘을 이용해 결측치를 대체하였습니다. 그리고 너무 많은 변수들도 Autoencoder를 통한 데이터 재구성에 악영향을 준다고 판단하여 RandomForest 기반의 알고리즘 VSURF를 이용해 일차적으로 유의미한 변수들을 선택하여 데이터의 변수들을 일차적으로 선별해 Autoencoder 의 단점 발생 가능성을 낮췄습니다.

이러한 AutoencoderClassifier 의 장점을 이용해 CASCADE 방법을 쓴다면, 구매 한 고객들에 대해서만 학습을 진행하여 좀 더 세분화 된 예측이 가능한 모형을 만들 수 있을 것이라 생각해 AutoencoderClassifier 와 회귀 MLP를 결합하는 방향으로 모형 구축의 방향을 결정했습니다.

3. 문제 해결 과정

앞서 설명한 착안점에 기반하여, 결측치 대체 부분과 변수선택 과정이 과연 유의미한 차이점을 발생시키는지 검증하기 위해 지도 분류용 Autoecndoer 모형을 구현하고, 결측치 대체를 평균과 최빈값으로 대체한 데이터와 MICE 알고리즘으로 대체한 데이터, 그리고 변수들을 VSURF를 이용해 일차적으로 선별한 데이터와 전체 변수를 그대로 이용한 데이터 총 4가지로 구분하여 Autoencoder 의 성능을 1차적으로 검증하였습니다.

그 결과 MICE를 이용해 결측치를 대체하고, VSURF를 이용해 학습시킨 모형이 가장 좋은 성능을 보여 이 데이터를 이용해 분석하는 것이 긍정적이라고 판단하였습니다. 선택한 데이터를 이용해 1차적으로 범주가 1 이상인 종속 변수들을 모두 1로 변환하여 train 데이터와 validation 데이터로 분리해 train 데이터로 지도 분류용 AutoencoderClassifier 모형을 이진 분류 모형으로 학습시켰습니다. 그후 검증 데이터를 같은 데이터를 학습한 LightGBM 과 XGBoost 모형과 결과를 비교하여 AutoencoderClassifier 의 성능이 더 좋은 결과를 보이는 것을 확인할 수 있었습니다.

○ 주요기술 (예선 계획서에서 사용한 알고리즘 구체화 또는 추가 알고리즘에 대해 자유롭게 기술

1) 결측치 대체 알고리즘 MICE : MICE 알고리즘은 여러 단계를 거쳐 결측치를 대체하는 알고리즘으로써, 단계별 여러 대체를 통해 결측치의 불확실성을 고려하여 단일 대체 방법에 비해 더 정확하고 안정된 추정값을 제공한다는 장점이 있습니다. MICE 는 chained 라고 부르는 주요 단계가 있는데 이는 다음과 같습니다.

① 초기 대체 : 각 결측치를 대체하기 위해 평균이나 중앙값과 같은 간단한 방법을 이용합니다.

② 순차적 대체 : 결측치가 있는 각 변수를 종속 변수로, 다른 변수들을 설명 변수로 사용하여 회귀 모형을 설정합니다. 그 후 해당 회귀 모형을 이용하여 결측치를 예측하고 대체합니다. 모든 변수에 대해 이 과정을 순차적으로 반복합니다.

③ 다중 대체 데이터 세트 생성 : MICE는 원본 데이터 세트에서 여러 개의 완전한 데이터셋을 생성합니다. 각 데이터 세트는 결측치 대체에 있어 약간의 차이를 띄고 있습니다.

④ 분석 & 결합 : 각 대체된 데이터 세트에 독립적으로 분석을 실시한 후, 얻은 결과들을 Rubin 규칙에 기반해 결합하여 하나의 결과를 얻습니다.

2) VSURF : 랜덤 포레스트를 기반으로 하는 변수 선택 방법으로써, 고차원 데이터에서 예측 성능을 유지하며 변수의 수를 줄이기 위해 유용한 변수만을 선택하기 위해 사용되는 알고리즘입니다. 높은 차원의 데이터에서 유용한 변수만을 선택하며 랜덤 포레스트의 예측 능력을 활용하여 변수 중요도를 비교적 정확하게 평가할 수 있으며, 다중 공선성 같은 문제에 강하다는 장점이 있습니다. VSURF는 크게 세 단계로 구성됩니다.

① 변수 필터링 단계 : 변수를 제거할 때 랜덤 포레스트 모델의 변동을 측정하여 모든 변수에 대해 랜덤 포레스트의 중요도를 계산합니다. 가장 중요도가 낮은 변수부터 순차적으로 제거하면서 랜덤 포레스트를 다시 학습시키고, 성능을 평가합니다. 이 과정을 반복하면서 모둔 변수에 대한 중요도 순위를 결정합니다.

② 변수 선택 단계 : out-of-bag 오류율을 최소화하는 지점을 임계값으로 설정하여, 중요도 순위를 기반으로 일정 임계값 이상의 중요도를 갖는 변수만을 선택합니다. 이 단계를 통해 불필요한 변수를 제거하고, 중요한 변수만을 선택합니다.

③ 해석 단계 : 선택된 변수 중에서도 예측에 크게 기여하는 변수와 데이터의 구조나

패턴을 설명하는데 도움이 되는 변수를 구분하여, 추가적인 랜덤 포레스트 모형을 사용하여 선택된 변수들의 중요도를 다시 평가하고, 두 가지 카테고리로 변수를 분류합니다.

3) 지도학습 분류용 Autoencoder 모형 : 기본적인 Autoencoder 모형은 비지도 학습 모형으로, 주로 차원 축소, 노이즈 제거 등의 작업에 사용됩니다. 입력 데이터를 압축된 표현으로 인코딩한 후, 이 표현을 다시 원래의 입력으로 디코딩하는 과정을 거치며 학습합니다. 이를 지도학습 분류용 AE 모형으로 구현할 경우 데이터의 복잡한 특성과 패턴을 자동으로 학습하며 , 재구성과 분류를 동시에 수행함으로써 모형이 주어진 데이터에 대해 더 깊은 이해를 얻을 수 있습니다. 모형의 기본 아이디어와 구조는 다음과 같습니다.

① 인코더와 디코더 : 기본적인 Autoencoder 와 마찬가지로 지도학습 분류용 Autoencoder 도 인코더와 디코더로 구성됩니다. 인코더는 입력 데이터를 압축된 표현으로 변환하고, 디코더는 이 압축된 표현을 다시 원본 데이터로 재구성합니다.

② 분류 계층 : 압축된 표현 위에 분류 계층을 추가합니다. 이 계층은 압축된 표현을 기반으로 데이터를 다양한 클래스로 분류하는 역할을 합니다.

③ 복합 손실 함수 : 모형은 재구성 오류와 분류 오류를 최소화하도록 이루어집니다. 전체 손실 함수는 이 두 손실의 조합으로 구성되어 이를 기반으로 훈련 데이터를 학습합니다.

4) 회귀용 MLP : MLP는 기본적으로 feedforward 인공 신경망의 한 종류입니다. MLP는 하나 이상의 은닉층을 포함하는데, 이는 MLP가 복잡한 패턴과 관계를 학습할 수 있게 해줍니다. 회귀 모형에 MLP를 이용할 경우, 복잡한 비선형 관계를 모델링 할 수 있고, 전통적인 회귀 모형으로 표현하기 어려운 데이터 패턴을 포착 가능하다는 장점이 있습니다. 또한, MLP의 은닉층의 수와 각 층의 뉴런 수 같은 구조를 조정 가능하므로 유연성이 높다는 장점이 있습니다. MLP의 구조는 다음과 같습니다.

① 입력층 : 특성의 수에 따라 뉴런의 수가 결정됩니다. ex) 10개의 특성을 가진 데이터의 경우, 입력층에는 10개의 뉴런이 있습니다.

② 은닉층 : 하나 이상의 은닉층을 포함할 수 있으며, 각 층은 다양한 수의 뉴런을 가질 수 있습니다. 은닉층과 뉴런의 수는 하이퍼파라미터로 이를 조정함으로써 모델의 복잡도를 조절할 수 있습니다.

③ 출력층 : 회귀 문제의 경우 일반적으로 출력층에 예측된 연속적인 값을 반환하는 하나의 뉴런만 존재합니다.

5) CASCADE 방법 : 이 방법의 기본 아이디어는 복잡한 분류 문제를 두 단계의 간단한 분류 문제와 회귀 문제로 나누는 것입니다. 첫 번째 분류 단계에서는 양성(전기차 구매함)과 음성(전기차 구매하지 않음)으로 분류됩니다. 음성으로 분류되면, 그 데이터는 후속 단계에서 더 이상 분석하지 않고, 양성으로 분류된 데이터들을 다음 단계의 회귀 모형으로 전달합니다. 이 방법을 이용하면 불균형 데이터로 인한 과적합 문제를 완화할 수 있고, 음성인 데이터를 거를 수 있기에 속도 향상의 장점이 있습니다. 구조는 다음과 같습니다.

① 분류단계 : 데이터 포인트가 두 그룹 중 특정 그룹에 속하는지 여부를 결정합니다. 구매 여부의 경우에는 구매한 그룹에 속하는지 여부를 결정하고, 구매 그룹에 속하는 데이터들을 다음의 회귀 단계로 보냅니다. 모델은 다양한 분류 알고리즘을 사용할 수 있지만, 이번 분석에서는 AutoencoderClassifier를 사용하였습니다.

② 회귀단계 : 이전 단계에서 구매 그룹에 속한다고 분류된 데이터 포인트들을 전달받아 회귀 분석을 수행합니다. 이는 특정 수치 값을 예측하는데 사용됩니다. 예를 들어 얼마나 전기차를 구매할 것인지를 예측하는 것과 같습니다.

○ 데이터 분석 및 모델링 과정

1. 학습용 데이터 가공

1) 결측치 처리

- 보수적인 관점으로 결측치가 5% 이상인 변수들 제거
- **나머지 결측치가 5% 미만으로 존재하는 변수들에 대해서는 두 가지 대체 방법 적용한 데이터 세트 2개 생성**
  (하나는 연속형은 평균, 범주형은 최빈값으로 결측치를 대체한 데이터 세트 A_1 와 나머지 하나는 MICE 알고리즘을 이용해 결측치를 대체한 데이터 세트 B_1 를 생성)

2) 이상치 처리

- CNT(전체 인구수) 보다 sum 관련 변수들 중 CNT 보다 큰 값이 존재하는 것을 발견
- 전체 인구수보다 가입자가 많은 것은 이상하다고 느꼈지만, 중복 가입이 존재할 수 있다고 판단하여 이상치 처리를 진행하지 않음

3) One-Hot encoding 및 변수선택

- 지역 변수에 대해 원-핫 인코딩을 적용
- BS_YR_Q 변수는 분기를 나타내는 시간 데이터이므로 이를 AutoencoderClassifier 에 반영하기 위해 순서형 범주로 변환 ex) 20201Q -> 0
- VSURF 방법을 이용하여 유의미한 변수들 선택
- 유의미한 변수들을 위의 A_1, B_1 데이터 세트에 적용하여 선택한 변수들만 존재하는 새로운 파생 데이터 세트 A_2(변수 선택된 결측치 평균,최빈값 대체 데이터 세트), B_2(변수 선택된 결측치 MICE 알고리즘 대체 데이터 세트) 생성

2. 모델 구축

1) 모델 선택

- 다중 분류 모형으로 AutoencoderClassifier 모델 선정
- 범주 0~12까지의 13개의 범주에 대한 다중 분류에 대한 모델을 구현하기로 선택

2) 하이퍼파라미터 튜닝

- 데이터를 target 변수의 범주 비율에 맞춰 train 과 test 데이터로 분할 (범주 12의 경우 데이터 개수가 1개이므로 제거 후 분할한 뒤 범주 12에 대한 데이터를 train 데이터에 삽입)
- LightGBM 과 XGBoost, AutoencoderClassifier 모델에 A_1, B_1, A_2, B_2의 train 데이터를 적용한 후, test 데이터를 이용해 accuracy 와 f1_score 값 비교
- AutoencoderClassifier 가 두 모델에 비해 좀 더 나은 성능을 보이지만 위의 평가값을 봤을 때 성능이 전체적으로 좋지 않다고 판단
- cross-validation을 이용해 하이퍼 파라미터 튜닝을 진행 후 다시 비교했지만 큰 차이가 없다고 판단

3) 대안 탐색

- AutoencoderClassifier의 활성화 뉴런에 대해 t-SNE 방법을 이용해 2차원으로 투영 후 시각화 했을 때, 범주 0과 관련된 뉴런과 범주 1~12 와 관련된 뉴런들이 잘 군집화 되었다는 것을 확인 -> 두 그룹 범주 0과 범주 1~12 까지의 데이터에 대해 뉴런이 데이터의 특성을 잘 학습했다고 판단
- 이를 이용해 CASECADE 방법을 이용해 1단계 : AutoencoderClassifier + 2단계 : MLP regression 으로 구성된 모형을 구축하기로 결정

3. 학습 방법

1) 1단계 AutoencoderClassifier 학습

- target 변수의 범주를 1~12까지를 1로 처리하여 0과 1의 값을 갖는 이진 변수로 변환
- CASCADE 모델에 범주가 0 (전기차 구매 X)과 1 (전기차 구매 O)로 변환된 데이터를 범주 비율에 맞춰 train 과 test 데이터를 분할 후 train 데이터를 학습
- 1단계 AutoencoderClassifier에서 0과 1의 그룹을 학습 -> 그 후 그룹 1에 속하는 데이터들만 추출
- 학습된 데이터에 대해 분류한 1의 개수가 4000개 정도이므로 전체 데이터 세트에서의 비율과 유사하다고 판단 -> 분류 모형이 잘 학습했다고 판단

2) 2단계 MLP Regression 학습

- 1단계에서 추출한 데이터의 target 변수를 '전기차 보유 수 / (국내차 보유 수 + 수입차 보유 수)' 로 변환 후 2단계 MLP Regression 적용
- 이후 test 데이터와 2단계에서 학습 후 예측한 predicition 데이터를 비교
- mse 값으로 예측 성능을 분석한 결과 좋은 성능을 발휘하는 것으로 판단.

○ 데이터 분석 및 모델링 방법론

1. 데이터 분석

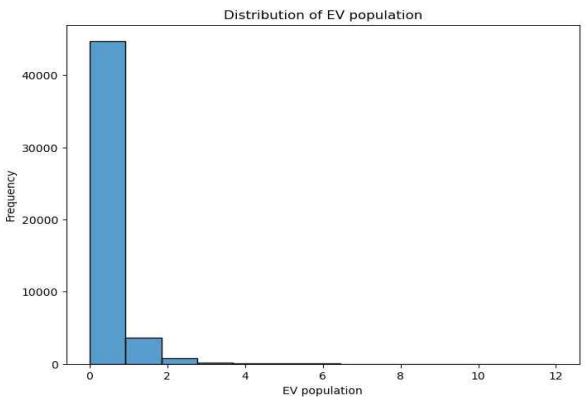
1) 데이터 분석

먼저 전체 데이터의 변수들 중 결측값을 가진 변수들이 무엇이 있는지 분석을 진행하였습니다. 총 213개의 변수 중에서 결측값이 하나라도 있는 변수는 119개이고 결측값이 10개, 100개, 1000개, 1000개, 20000개가 넘는 변수의 수는 각각 112개, 97개, 67개, 29개, 27개로, 결측값이 1000개가 넘는 변수의 수와 10000개가 넘는 변수의 수는 차이가 컸지만 결측값이 10000개가 넘는 변수의 수와 20000개가 넘는 변수의 수의 차이는 2개로 작았습니다. 또 결측값이 5% 이하인 변수의 개수와 결측값이 10% 이하인 변수의 개수가 큰 차이가 없었기 때문에 전처리 과정에서 더 엄격한 임계값인 5% 를 기준으로 5% 이상의 결측치를 가진 변수들을 제거하였습니다.

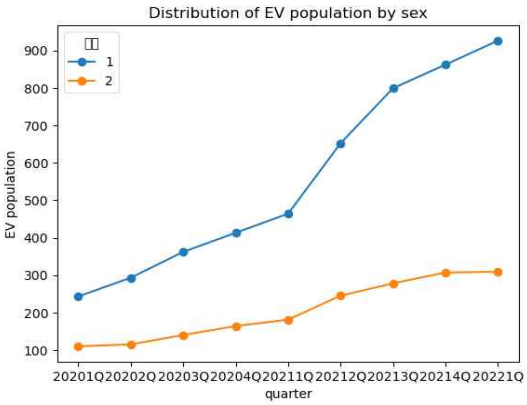
이후 Target 변수에 대해서도 분석을 진행하였습니다. Target 변수는 0부터 12까지의 범주를 갖는 데이터로 판단하였고, 각 범주들이 차지하고 있는 비율을 분석한 결과 범주가 0인 데이터가 90 % 이상일 정도로 0을 제외한 나머지 범주들의 데이터의 개수가 매우 불균형한 데이터였습니다. SMOTE를 이용한 언더샘플링이나 오버샘플링의 경우, 과적합이나 정보 감소 같은 치명적인 단점들이 존재하기에 이에 대한 처리방법을 반영하여 모형 구축 계획을 세우기로 결정하였습니다.

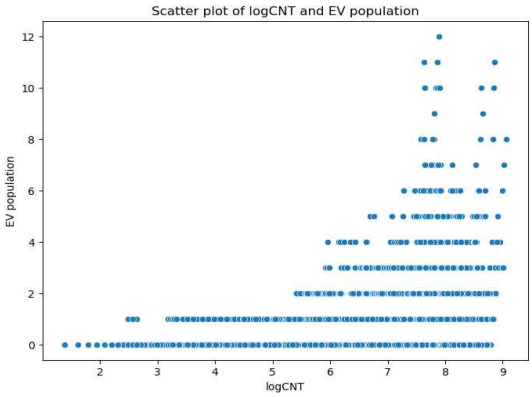
2) EDA

우선 첫 번째로, Target 데이터인 전기차 보유 인구수의 분포를 EDA를 통해 시각적으로 구현하였습니다. 밑의 그림처럼 skewed한 분포를 띄고 있는 불균형 데이터라는 것을 다시 한번 확인할 수 있었습니다. 이러한 데이터를 기반으로 모형을 학습할 시 데이터가 많은 범주에 과적합될 가능성이 높아 매우 고민이 많아졌습니다.

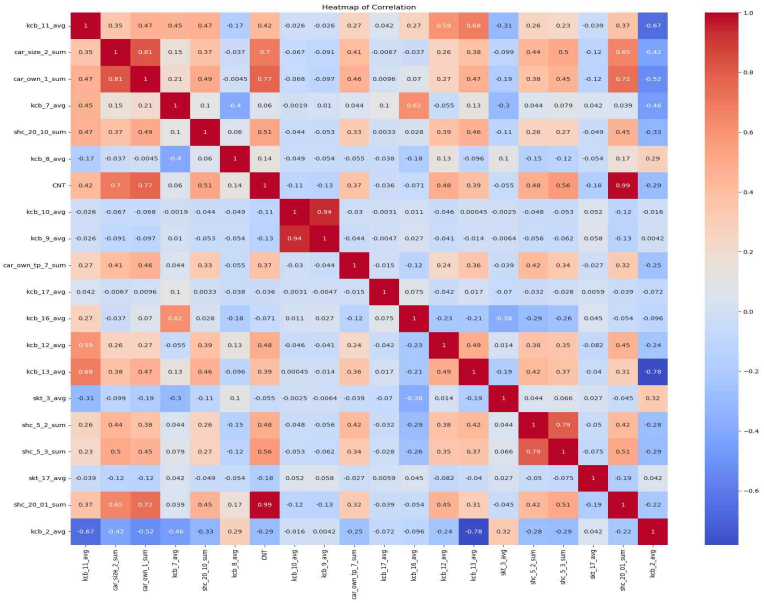


두 번째로, 시간 변수인 'BS_YR_Q' 와 'CNT'에 대하여 전기차 보유 인구수가 어떻게 변화하는지 관찰하였습니다. 시간에 따른 전기차 보유 인구수의 경우, 이를 성별로 구분하여 시각화 하였습니다. CNT 의 경우에는 숫자가 너무 커 그래프의 경향성이 파악하기 쉽지 않아 로그 변환을 진행하였습니다. 확인 결과, 밑의 그림처럼 시간이 증가함에 따라 전기차 보유 인구수가 증가하였고, 전체 인구수가 많을수록 전기차 보유 인구수가 많아 진다는 것을 확인할 수 있었습니다. 이를 통해 시간이 지남에 따라 전기차에 대한 대중들의 인식이 조금씩 긍정적으로 변해갔다는 인사이트와 남성이 여성보다 전기차 구매에 긍정적이라는 인사이트를 도출하였습니다.





마지막으로 LightGBM을 통해 최중요 변수 20개를 선정하여 각 변수들 간의 상관관계를 분석했습니다. 분석한 결과 '추정연소득 평균'과 '신용대출잔액 평균'이 높은 음의 상관관계를 보였습니다. 또, '신한 최근 1년 월평균 이용금액 100만원이하 인구수'와 '집계 인구수', '국산차보유 인구수'와 '보유차량(소형) 인구수'가 각각 상관계수 0.99, 0.81로 높은 양의 상관관계를 보였습니다.



3) 전처리

다소 보수적인 관점으로 5% 이상의 결측치를 가진 열을 삭제함으로써 일부 결측치를 가지는 변수들을 처리하였습니다. 많은 결측치를 대체할 경우 데이터의 크기는 커질 수 있으나 오히려 데이터가 왜곡되어 예측 성능을 저하시킬 것을 우려하였기 때문입니다. 결측치에 대해 엄격한 기준을 세우는 대신 분석방법과 변수선택에 더 집중하는 것이 모델의 예측 성능과 해석력을 높일 수 있다고 판단하였습니다. 결측치를 대체하기 위해 연속형 변수는 평균 대체, 범주형 변수는 최빈값 대체로 진행하려 하였지만, 너무 단순한 대체는 불균형 데이터에서 모형 생성에 악영향을 줄 것이라 판단하였고, 더 좋은 방법을 모색하였습니다. 그러다 MICE 알고리즘을 떠올렸고, 이 알고리즘을 통해 나머지 5% 미만의 결측치를 가지는 변수들에 대해서는 관련된 정보를 최대한 반영함으로써 결측치 대체로 인한 데이터 왜곡 위험을 줄이고자 하였습니다.

시군구코드는 분석개체 집단을 구별 짓는 핵심 변수로 판단하여, 데이터의 차원이 커지는 것을 감수하더라도 one-hot encoding을 통해 독립변수로 설정하였습니다. 그리고 데이터 탐색 결과 전기차 보유 인구수가 시간의 흐름에 따라 점점 증가하는 추세를 보였기에 기준시점을 나타내는 'BS_YR_Q' 변수를 머신러닝과 딥러닝 모형들에 원활히 사용하기 위해 2020년 1분기를 0 이라는 기준으로 놓고 순서형 범주로 변환하였습니다.

4) 변수 선택

전처리 과정까지 마친 상태에서 기존의 통계학적 상식에 근거해 데이터의 변수가 데이터의 크기에 비해 많아 분석 과정에서 차원의 저주에 빠질 수도 있다고 판단하였고, 변수 선택 과정을 거치기로 하였습니다. 변수 선택 방법으로는 VSURF 방법을 이용했습니다. 먼저 target의 범주 1~12 까지를 1로 변환하여 traget 데이터를 구매를 했는지 안했는지에 대한 이진 데이터로 변환하였습니다. 이 target 데이터에 대해 독립 변수들을 VSURF 알고리즘에 학습시켰습니다. 그리고 다시 target 데이터를 주최 측에서 정해준 회귀 변수로 변환하여 VSURF 알고리즘에 학습을 진행하였습니다. 그 후, 두 방법에서 나온 중요 변수들의 교집합을 최종 독립 변수들로 설정하였습니다.

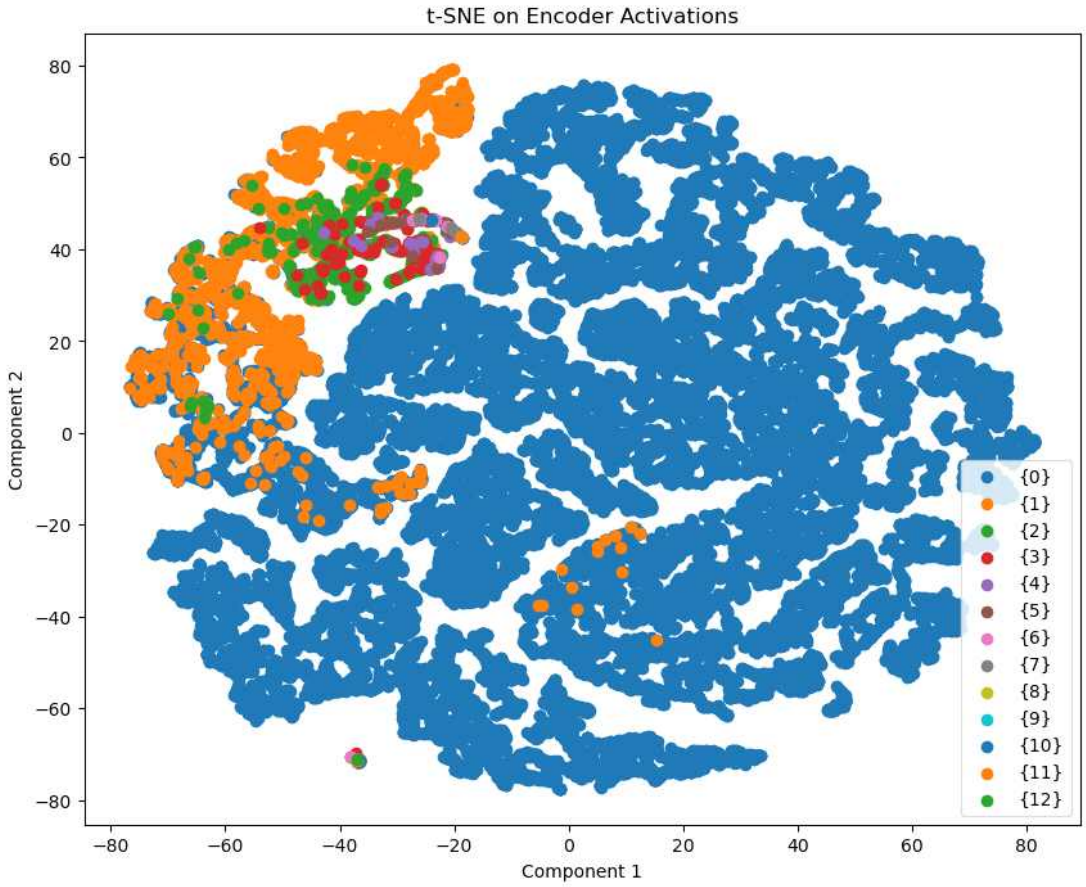
저희 팀은 모델 구축에 대해 분류와 회귀를 합친 CASCADE 방법을 이용할 계획을 세웠기에 위와 같은 변수 선택 방법을 진행하여 분류와 회귀 방법 둘 다에 대해 의미있는 변수들을 추출하였습니다.

4. 모델 구축

1) 모델 선택

우선 처음에는 다중 분류 모형을 고려하여 AutoencoderClassifier를 분석 모형으로 선정할 생각이었습니다. 다른 머신러닝 분류 모형들과 똑같이 학습한 후, 각 모형을 통해 비교하여 AutoencoderClassifier 모형의 성능의 우수성을 보이며 예측 모형으로 사용할 계획을 가지고 있었습니다. 하지만 Autoencoder 의 성능이 다른 모형들보다 조금 더 높게 나왔지만, 모든 분류 모형들이 만족할만한 성능을 보이지 않아 분석의 방향성을 팀원들과 다시금 고민했습니다.

팀원 중 한 명이 CASCADE 방법론을 제시하며, 이에 대한 효과와 현재의 불균형에 적합한 당위성을 제시하여, 기존의 Autoencoder 와 MLP를 섞어 CASCADE 방법론을 구현하기로 결정했습니다. 이 같은 결정을 내린 근거에는 Autoencoder 의 활성화 뉴런들에 대해 t-SNE 을 이용하여 2차원 시각화를 진행해본 결과, 범주 0과 범주 1~12 에 대해 뉴런들이 군집화하는 형태를 띄고 있어 범주에 맞는 학습을 했다는 것을 알 수 있습니다. 시각화의 그림은 다음과 같습니다.



보시다시피 0과 1~12 까지의 범주가 군집화 되어 있는 것을 알 수 있습니다. 따라서 앞서 주요기술에서 설명한 CASCADE 방법에 분류 모형으로 Autoencoder를 이용하는 것이 적절하다는 결론을 내리고 모형 구축을 진행하였습니다.

2) 학습 방법

먼저 데이터를 분할하기 위해 train과 test를 나눌 때 train의 class 비율을 유지하면서 test가 split 되도록 하였습니다. 또 데이터가 0의 비율이 매우 높으므로 학습 시 0이 아닌 값에 대해 더 많은 가중치를 설정하여 효과적인 학습을 하도록 하였습니다.

CASECADE 모델의 첫 번째 단계인 이진 분류에서 학습할 때는 train set을 사용하였으며 1 이상의 값은 모두 1로 코딩하여 이진분류 문제로 학습하였습니다. 그리고 1 의 값으로 예측되는 행들의 인덱스를 추출하여 1로 예측된 데이터들을 1차적으로 추출하였습니다. 두 번째 단계인 회귀에서는 1단계에서 추출된 train 데이터들을 이용하여 MLP Regression을 학습하였습니다.

3. 검증

회귀용 target 변수로, target 변수를 변형한 test 데이터를 이용해 mse 값을 비교해 봤더니 매우 낮은 수치를 나태내어 성능이 좋은 모형을 학습했다고 판단했습니다. 시간이 부족해 하이퍼 파라미터 튜닝을 하지 못해 아쉬움이 남지만, 지금 그 자체로도 충분히 좋은 모델이라 판단해 그대로 사용하기로 결정했습니다. 최종 모델의 경우 보유한 데이터 전체를 train set으로 간주하여 학습하였고, 이 모델을 저장 후 제출하였습니다.

○ 최종결과

1. 결과 요약

VSURF를 이용해 상위 10개 정도의 변수를 추출해봤더니 'kcb_12_avg'(주택담보대출잔액 평균), 'skt_35_avg'(쇼핑 소셜커머스 이용일수 평균), 'skt_48_avg'(엔터 사진 이용일수 평균), 'kcb_13_avg'(신용대출잔액 평균), 'kcb_43_avg'(카드업종대출잔액 평균), 'shc_12_3_sum'(신한 게임이용등급), 'skt_51_avg'(엔터 영화 이용일수 평균), 'home_sigun_44150()', 'skt_55_avg'(위치 교통 이용일수 평균), 'shc_21_avg'(요식 이용금액 평균) 의 변수들이 전기차 구매에 큰 영향을 미치는 것으로 파악됩니다.

CASCADE 모델을 이용해 분석한 결과 train 단계에서는 1단계 분류 모형의 auc 가 0.88 정도가 나오고 2단계 회귀 모형의 mse 가 0.0001로 나왔습니다. test 단계에서는 1단계 test auc 가 0.82이고 2단계 test mse가 0.0001 로 나왔으므로 매우 준수한 성능을 내는 모형이라 판단하였습니다.

2. 참가팀의 핵심 아이디어

CASCADE 모델이 핵심 아이디어였다고 생각합니다. Autoencoder 와 MLP 각각의 모델도 참신한 분석 접근 방법이었지만, 불균형한 데이터로 좋은 성능을 내지 못하는 상황에서 두 모델의 장점들 강화하여 단계별 분석을 통해 각각의 모델로는 낼 수 없는 성능을 발휘하는 모델을 구현할 수 있었다는 것이 저희 참가팀의 핵심 아이디어라고 생각합니다.

3. 기대효과

1) 높은 예측 정확도

MLP와 Autoencoder를 조합한 CASCADE 모델을 통해 데이터의 불균형 문제를 극복하고 추출된 중요 변수를 기반으로 이전보다 더 정확한 예측을 기대할 수 있습니다.

2) 변수 중요도 파악

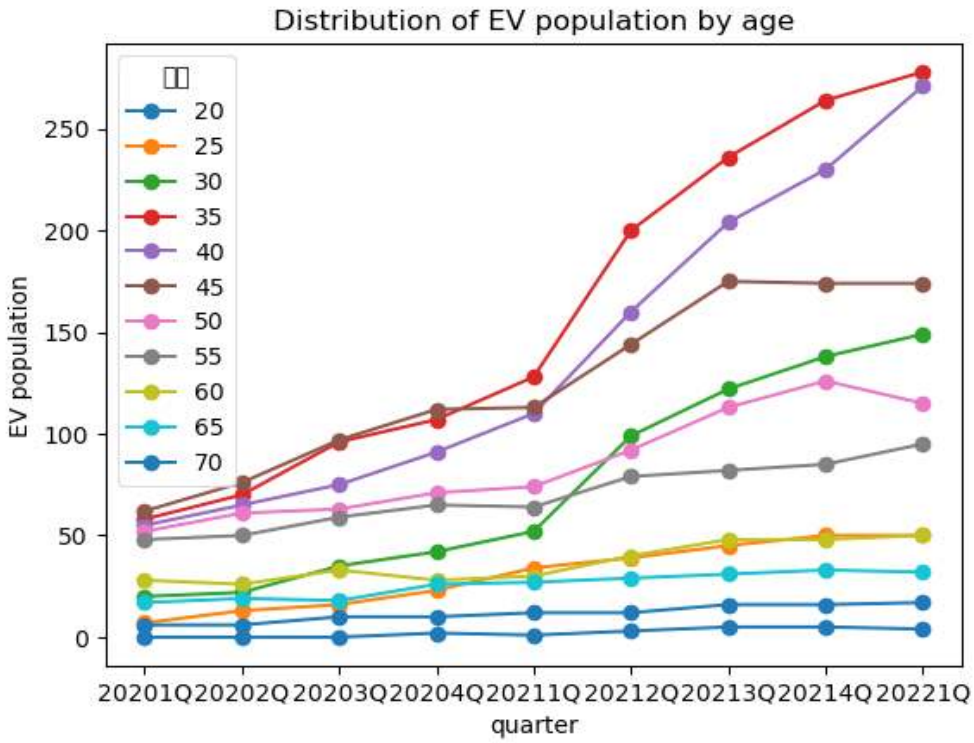
LightGBM의 변수 중요도와 VSURF를 통해 전기차 구매 예측에 영향을 미치는 주요 변수들을 파악할 수 있습니다. 이 정보를 통해 더 효과적인 마케팅 전략과 판매전략을 설계할 수 있습니다. 예를 들어, LightGBM의 이진 분류에서 대출잔액과 소득과 같이 경제적 수준을 나타

내는 변수가 중요한 변수로 선택된 것으로 보아 경제적 여유가 있는 집단을 상대로 마케팅 전략을 수립하는 것이 유리할 것으로 예상됩니다.

3) 전략적 판매계획 수립

전기차 보유 예측 모델을 활용하여 효과적인 판매전략을 계획할 수 있습니다. 특정 고객 그룹의 판매 가능성을 더 정확히 파악하면, 해당 그룹에 맞춤형 마케팅 전략을 수립할 수 있습니다.

가령, 20대보다 30대 후반 ~ 40대 후반의 전기차 보유 인구수가 많은 것을 확인할 수 있었는데, 해당 연령층은 사회초년생인 20대보다 경제적 활동이 활발하여 소득이 상대적으로 높기 때문인 것으로 예상됩니다. 또 30대 후반 ~ 40대 후반 내에서는 연령층이 낮을수록 전기차 보유 인구수가 많으므로 트렌드에 민감한 낮은 연령층이 차세대 전기차 고객이 될 것임을 예상할 수 있습니다. 따라서 잠재적 전기차 고객인 젊은 연령층의 트렌드를 고려하면서도 청장년층의 전기차에 대한 구체적인 수요 요인을 정확하게 파악하는 것이 중요하다고 할 수 있습니다.
또, 전기차를 1대도 보유하지 않은 집단이 전체인구의 약 90%이므로 기존 고객 유지보다는 새로운 고객 유입을 촉진하기 위해 강력한 마케팅 전략을 구사하는 것이 권장됩니다.



4) 판매 예측 및 재고 관리

전기차 예상 수요를 예측하여 효율적인 판매계획을 세울 수 있습니다. 이를 토대로 재고 관리를 최적화하여 비용을 절감하고 효율성을 높일 수 있습니다.

5) 개인화된 고객 서비스

고객 예측 모델을 활용하여 전기차 구매 가능성이 높은 개별 고객에게 맞춤형 서비스를 제공할 수 있습니다. 이를 통해 고객 만족도를 향상시키고, 더 나은 고객 경험을 제공할 수 있습니다. 또한,

종합적으로 이 프로젝트는 전기차 보유 예측 모델을 통해 전기차 시장 고객에 대한 깊은 이해를 제공하고, 이를 바탕으로 마케팅 전략을 구체화하여 마케팅 비용을 절감하고 효과적인 판매전략을 수립하는 데 기여할 것으로 기대합니다.

4. 공모전 소감

이 프로젝트를 참여하며, 시간적 여유가 부족하여 완벽한 분석 과정을 진행하지 못했다는 점이 팀원들 전체적으로 아쉬움이 남지만, 여러 통계, 머신러닝, 딥러닝들에 관한 새로운 방법론들을 찾아보고 이를 적용해보며 책으로 배운 지식으로만 남아있던 것들을 직접 체화하고, 새로 더해가며 발전시켜 팀원 전체의 성장을 이뤄낼 수 있어 좋은 경험이었습니다.

불균형 데이터에 대해 분석할 수 있는 여러 방법론과 논문들이 있었지만, 데이터의 종류나 실제 데이터에 따라 큰 효과를 얻지 못하는 방법론들이 많았습니다. 이 같은 것들을 직접 시험해보고 선별하여 분석에 적용할 수 있었던 것도 좋은 경험이었다고 생각합니다.

가장 크게 느꼈던 점은 이미지 분류에 더 적합한 딥러닝 모델들을 Tabular 데이터 적용하여도 좋은 성능을 내는 모델을 만들 수 있었다는 점입니다. 이런 방법들이 점차 개발되면 우리가 가지고 있는 일반 데이터들에서도 더 좋은 성능을 내는 딥러닝 모델들을 만들 수 있다는 것이 이번 분석의 의의가 아니었나 싶습니다.