별첨2 **데이터 분석·모델링 개발 계획서(안)**

□ 개요

1. 요약

본 분석에서는 전기차 구매 예측 모델을 개발하며, 주로 MLP와 트리 기반 알고리즘을 사용하여 전기차 구매 예측 및 주요 변수들에 대한 인사이트를 도출합니다. 또한 다른 방식으로 딥러닝 기반 클러스터링을 통해 그룹화되어, 각 그룹의 전기차 보유율과 관련된 특성을 분석합니다. 이를 통해 전기차 구매 경향에 대한 깊이 있는 이해를 도출하고, 효과적인 마케팅 전략 및 판매 계획을 수립할 수 있을 것입니다. 이러한 과정은 판매 예측, 재고 관리, 고객 서비스 개선 등 다양한 실질적 효과를 가져올 것으로 기대됩니다.

2. 주요 방법론

① 본 분석에서는 전기차 보유 여부를 타겟 변수로 지정해 지도학습을 실시하여 고객이 전기차를 구매할 것인지에 대해 예측하는 모델을 개발하고 어떤 피처 변수들이 전기차 구매에 영향을 미치는 지 파악하여 전기차 구매 고객의 특성에 대한 인사이트를 도출하는 것이 목표입니다.

이를 위해 주요 방법론으로 딥러닝의 한 형태인 Multi-Layer Perceptron (MLP)를 중점적으로 활용하며, 부가적으로 LightGBM, Random Forest, CatBoost와 같은 트리 기반 알고리즘을 참고 모형으로 활용할여 예측 성능과 변수 중요도를 비교할 예정입니다.

MLP 모델에는 Penalized Neural Network 방법론을 사용하여 의미없는 변수들을 제거하고, 연관성이 있는 변수들을 그룹화하여 올바른 설명력을 갖는 모형을 만들 예정입니다.

MLP는 비선형 패턴을 잘 학습할 수 있는 장점이 있으나, 동시에 모델의 해석력이 상대적으로 낮은 편입니다. 이에 반해, 트리 기반 모형들은 뛰어난 해석력을 가지고 있으며, 중요 변수의 선정에 있어 유용한 정보를 제공합니다. 따라서, MLP 모델의 예측 성능을 확보하는 한편, 트리 기반 모형들을 통해 변수 중요도를 파악하고 비교하는 방식으로 모델링을 진행할 계획입니다. 이를 통해, 각 모델들의 장점을 최대한 살린 통합적인 분석 방법론을 구축하게 됩니다.

② 이 프로젝트의 또 다른 접근 방식은 딥러닝 기반 클러스터링을 활용하여 고객 데이터를 분석하는 것입니다. 초기 단계에서는 고객 데이터를 비지도 학습 방법인 딥러닝 클러스터링 기법을 사용하여 서로 다른 특성을 가진 3~5개의 그룹으로 세분화합니다.

그리고 고급 클러스터링 알고리즘 중 하나인 Autoencoder를 활용할 예정입니다. Autoencoder는 데이터의 복잡한 패턴을 잘 학습하며, 데이터의 중요 특성을 자동으로 추출하여 차원을 축소하는 능력을 가진 딥러닝 모델입니다. 이를 통해 데이터의 복잡도를 줄이고 중요한 특성을 강조할 수 있습니다.

이렇게 클러스터링 된 데이터는 각 그룹별로 '전기차 보유 여부'의 비율을 측정함으로써, 특정 특성을 가진 고객 그룹이 전기차를 많이 보유하고 있는지를 판단합니다. 이런 방식으로 각 고객 그룹의 특성을 파악하고, 그룹 간 전기차 보유율의 차이를 이해하는 데 도움을 줍니다.

그 다음 단계에서는 이전에 언급한 MLP, LightGBM, Random Forest, CatBoost와 같은 모델을 각각의 클러스터에 적용하여, 클러스터마다 어떤 특성이 전기차 보유에 가장 큰 영향을 미치는지를 확인합니다. 이를 통해 클러스터마다 다르게 작용하는 특성을 파악하고, 이를 바탕으로 더욱 효과적인 마케팅 전략을 개발할 수 있습니다.

이렇게 복합적인 방법론을 적용함으로써 고객의 전기차 구매 경향에 대한 깊이 있는 이해를 얻고, 이를 바탕으로 실제 비즈니스 전략에 효과적으로 활용할 수 있습니다.

3. 분석 • 모델링 기법 선택 배경

MLP와 트리 기반 모형들의 조합은 다양한 수치형과 범주형 변수들, 비선형 패턴, 복잡한 상호작용 등의 데이터셋 특성과 잘 맞습니다. MLP는 비선형 패턴을 학습하고 표현하는 능력, 높은 유연성, 자동 특성 학습, 대용량 데이터 처리 능력 등의 이점 때문에 선택되었으며, 트리 기반 모형들은 각 변수의 중요성을 명확하게 이해하고 해석할 수 있는 능력 때문에 참고 모형으로 선택되었습니다.

하지만 MLP 모델은 변수 선택의 효과가 존재하지 않아, 모든 변수를 포함하여 모델을 만든다는 단점이 존재하므로 "Variable Selection via Penalized Neural Network: a Drop-Out-One Loss Approach" 와 "Sparse-Input Neural Networks for High-dimensional Nonparametric Regression and Classification" 라는 논문을 참조하여 Sparse 한 MLP 모델을 구현하여 변수 선택의 효과를 추가해 더 나은 분류 예측 모델을 만들 수 있을 것으로 기대합니다.

모델링 접근법은 데이터의 특성과 문제 설정에 따라 크게 달라집니다. 따라서 이 프로젝트에서는 전기차 보유고객의 특성과 전기차 보유율의 차이를 파악하기 위해 비지도 학습 방법인 딥러닝 클러스터링 기법을 사용하여 추가적인 분석을 진행할 예정입니다. 고객의 세분화를 위해 Autoencoder라는 고급 클러스터링 알고리즘을 활용할 생각입니다.

4. 기대효과

MLP와 트리 기반 모형들의 조합을 이용한 모델링을 통해, 다음과 같은 기대효과가 있습니다

높은 예측 정확도: MLP를 통한 비선형 패턴 학습과 트리 기반 모형들의 변수 중요도 분석을 통해 전기차 보유 여부를 더 정확하게 예측할 수 있습니다.

변수 중요도 파악: 트리 기반 모형들의 결과를 통해 어떤 변수가 전기차 구매 예측에 큰 영향을 미치는지 파악할 수 있습니다. 이를 통해 마케팅 전략이나 판매 전략을 보다 효과적으로 개발하는데 도움이 될 것입니다.

전략적 판매계획 수립: 예측 모델을 통해 얻은 인사이트를 활용하여 전기차 판매에 대한 효과적인 전략을 개발할 수 있습니다. 예를 들어, 특정 소비자 그룹이 전기차를 구매할 가능성이 높다는 것을 파악한다면, 해당 그룹을 대상으로 한 마케팅 전략을 수립할 수 있습니다.

데이터 이해도 증진: 복잡한 패턴과 상호작용을 학습하고 표현할 수 있는 MLP 모델을 통해 전체 데이터에 대한 이해도를 높일 수 있습니다. 이는 기업의 전기차 판매 전략을 개선하고 시장 확대에 기여하는데 중요한 역할을 할 것입니다.

판매 예측 및 재고 관리: 전기차 예상 수요고객을 파악함으로써 판매량을 예측할 수 있습니다. 이에 따라 효율적인 재고 관리를 할 수 있습니다. 효율적인 재고관리는 기업의 효과적인 비용 절감에 도움이 될 것으로 기대합니다.

개인화된 고객 서비스: 고객 예측 모델을 활용하여 개인화된 고객 서비스를 제공할 수 있습니다. 예상 고객의 취향, 구매력 등을 파악하여 맞춤형 추천, 서비스 개선 등을 통해 고객 만족도를 향상시킬 수 있습니다.

총괄적으로, 본 프로젝트는 전기차 보유 예측 모델을 통해 전기차 시장의 고객에 대한 깊은 이해를 제공하고, 이를 바탕으로 마케팅 전략을 개발하는 토대로 삼아 마케팅 비용을 줄이고 효과적인 판매 전략을 세우는데 도움을 줄 것으로 기대합니다.

□ 참가팀의 핵심 기술 설명

핵심 기술: Penalized Neural Network

2018년과 2019년에 발표된 논문을 기반으로한 Penalized Neural Network 모델링 기법은 통계학에서 사용되는 Sparse Method와 Multi-Layer Perceptron(MLP)를 결합한 방법입니다. 이 방법은 데이터 분석과 예측 모델링에 있어서 변수 선택과 모델 복잡성 조절을 동시에 수행할 수 있는 강력한 도구로 간주됩니다.

Sparse Method는 통계학에서 변수 선택에 활용되는 방법론으로, 주어진 데이터에서 유의미한 변수만을 선택하여 모델을 구축하는 것을 목표로 합니다. 이를 통해 모델의 복잡성을 줄이고 해석 가능성을 높일 수 있습니다. Sparse Method는 변수 선택에 대한 제약 조건을 추가하여 변수의 개수를 제한하는 방식으로 작동합니다. 주로 L1 regularization이나 Lasso regression과 같은 기법이 사용됩니다.

반면, Multi-Layer Perceptron(MLP)은 인공신경망(ANN)의 일종으로, 여러 개의 은닉층을 가진 신경망 구조입니다. MLP는 비선형 관계를 모델링하고 복잡한 패턴을 학습할 수 있어 다양한 분야에서 활용되고 있습니다. MLP는 변수 간의 비선형 관계를 모델링하고, 복잡한 패턴을 학습하여 예측 성능을 향상시킬 수 있습니다.

Penalized Neural Network 모델링 기법은 이러한 Sparse Method와 MLP를 결합하여 변수 선택과 모델 복잡성 조절을 동시에 수행합니다. 이를 통해 예측 모델의 성능을 향상시키고, 중요한 변수들에 대한 해석 가능성을 제공할 수 있습니다. 변수 선택을 통해 불필요한 변수를 제거하고 모델의 복잡성을 줄이는 동시에, MLP를 사용하여 비선형 관계와 복잡한 패턴을 모델링할 수 있습니다.

Penalized Neural Network 모델링은 다양한 분야에서 활용될 수 있으며, 변수 선택과 모델 복잡성 조절의 중요성이 강조되는 데이터 분석 및 예측 문제에 적합한 방법입니다.

아래는 Penalized Neural Network 모델링 기법의 주요 학습 단계를 정리한 것입니다:

1. 초기화: Xavier 메소드를 사용하여 가중치를 초기화합니다. Xavier 초기화는 신경망의 각 레이어에서 가중치를 적절하게 초기화하기 위한 방법으로 널리 사용됩니다.

2. 매개변수 업데이트: Gradient Descent 방법을 사용하여 매개변수를 업데이트합니다. Gradient Descent는 손실 함수의 기울기를 따라 매개변수를 조금씩 업데이트하여 최적의 솔루션을 찾는 최적화 알고리즘입니다.

3. Lasso 및 Group Lasso Shrinkage: 첫 번째 레이어의 매개변수에 대해 Lasso 및 Group Lasso shrinkage를 적용하기 위해 소프트 쓰레스홀딩(soft-thresholding) 연산자를 사용합니다. 이를 통해 변수 선택과 모델의 복잡성을 조절할 수 있습니다.

4. Line Search Criterion 확인: Line Search Criterion을 만족하는지 확인합니다. Line Search Criterion은 매개변수 업데이트 후 손실 함수의 감소 정도를 측정하여 적절한 학습 속도를 결정하는 기준입니다.

5. Line Search Criterion 충족 여부에 따른 처리: 만약 Line Search Criterion을 만족한다면, 이전 단계에서 적용한 매개변수 업데이트를 유지하고 다음 반복으로 진행합니다.

6. Line Search Criterion 불충족 시 재학습: Line Search Criterion을 만족하지 않는다면, 매개변수를 이전 상태로 되돌리고 learning rate decay 방법을 사용하여 학습률을 감소시키고 다시 재학습을 시도합니다. 이 과정은 모델이 수렴할 때까지 반복됩니다.

위의 단계들을 통해 Penalized Neural Network 모델링 기법은 변수 선택과 모델 복잡성 조절을 동시에 수행하며, 최적의 모델을 찾을 수 있도록 합니다.

□ PoC(Proof of Concept) 프로그램 설명

ㅇ 데이터 명세서와 샘플 기반으로 만든 PoC 프로그램

```
# model structure

class SPINN(nn.Module):
    def __init__(self, input_dim, h1_dim, h2_dim):
        super(SPINN, self).__init__()
        self.input_dim = input_dim
        self.hidden_dim1 = h1_dim
        self.hidden_dim2 = h2_dim

        # in this case solve regression problem

        self.layer1 = nn.Linear(self.input_dim, self.hidden_dim1)
        self.layer2 = nn.Linear(self.hidden_dim1, self.hidden_dim2)
        self.layer3 = nn.Linear(self.hidden_dim2, self.hidden_dim2)
        self.layer4 = nn.Linear(self.hidden_dim2, 1)

    def forward(self, X):

        # activation function

        out_h1 = torch.tanh(self.layer1(X))
        out_h2 = torch.relu(self.layer2(out_h1))
        out_h3 = torch.tanh(self.layer3(out_h2))
        final_out = self.layer4(out_h3)

        return final_out
```

```
############################## update rule ######################

# define optimizer

optimizer = optim.SGD(model.parameters(), lr = learning_rate)

# Gradient Descent algorithm

optimizer.zero_grad()
loss.backward()
optimizer.step()


# Group Lasso update using self define function

model.state_dict()['layer2.weight']                                    =
            GroupLasso_update_p(model.state_dict()['layer2.weight'],lambda_1,
            optimizer.param_groups[0]['lr'])

# calculate difference of parameter between previous parameters and updated
            parameters using self define function
param_diff = l2_loss(current_params, best_params)

# Line search criterion
if train_custom_list[epoch] < best_loss - param_diff:
    best_loss = train_custom_list[epoch]
    # save updated parameter
    best_params = copy.deepcopy(current_params)
else:
    # load previous parameters
    model.load_state_dict(best_params)
    # learning rate decay
    optimizer.param_groups[0]['lr'] *= 0.9
```

파이썬 기반의 Group lasso 방식을 이용한 MLP 코드이며, 주석은 '#영어' 로 작성하였습니다.

ㅇ 개발 도구

활용하는 도구(프로그램, 라이브러리) 또는 기술

구분	도구 이름	버전	제조사(출처)	용도
1	Torch	1.12	FAIR(Facebook AI Research)	
2	Scikitlearn			
3	Numpy, Pandas			

□ 예상 결과

LightGBM, Random Forest, CatBoost와 같은 트리 기반 모델 및 군집화 분석을 통해 전기차 예상 수요고객의 특성에 해당하는 피쳐를 추출함으로써 전기차 수요고객의 특성에 대한 인사이트 도출을 기대합니다. 가령 전기차 이용이 용이해야하므로 전기차 충전소가 많은 지역에 속한 사람일수록, 자택과 직장에서 충전소까지의 최단거리가 짧을수록 전기차를 구매할 것으로 예상됩니다. 또, 새로운 기술과 혁신에 관심이 있는 사람들이 전기차를 구매하는 경향이 있으므로 기술트렌드에 민감한 연령층이나 관련 활동을 하고 있는 사람일수록 전기차를 구매할 것으로 예상됩니다.

트리모델 및 군집화로 필터링 된 피쳐들을 기반으로 Multi-Layer Perceptron(MLP)를 시행하여 전기차 예상 수요고객을 예측할 수 있을 것으로 기대합니다.