Data Mining Term Project

Bank Marketing Campaign

2018250028
Biomedical Engineering
Younsuk Yeom
염윤석

1. Introduction

은행 고객 정보, 사회 및 경제 현황 정보 및 이전 분기 마케팅 활동 정보 등이 담겨있는 데이터를 바탕으로 고객들의 다음 분기 예금 상품 등록 여부를 예측하고, 이를 통해서 예금 상품 마케팅 전략을 세워보고자 한다.

사용한 데이터는 UCI archive[1]의 Bank Marketing 데이터[2]로 포르투갈 금융 기관의 다이렉트 홍보 관련 데이터이다. 전화 채널을 통한 마케팅 데이터이며, 같은 고객에 대해 적어도 한번 이상의 연락을 취해 얻은 데이터이다. 해당 고객들의 기본 인적 사항 뿐만 아니라, 현재 경제 지표, 이전 분기 마케팅 활동 정보들을 포함하여, 해당 고객이 정기 예금 예치 여부를 조사하여 수집한 데이터셋이다. 데이터는 2008년 5월부터 2010년 11월까지 수집되었으며, 총 41188건의 데이터로 이루어져 있다. 1개의 binary 종속 변수를 제외한, 20개의 독립 변수들로 구성되어 있으며, 독립변수는 크게 세가지로 나뉜다.[Table 1]

전체 데이터 마이닝 작업 흐름은 다음과 같다. 20개의 변수 구성 및 분포를 확인하여, 변수 간의 관계를 파악하고, 예금 상품 등록 여부(Target Variable)를 중점으로 가설들을 수립한다. 이후, EDA(Exploratory Data Analysis)를 통해서, 수립된 가설들을 검증하는 과정을 거쳐, 학습 모델 형성을 위한 중요 변수들을 선택한다. 수치형 변수에 대해서는 Scaling을 범주형 변수에 대해서는, One hot Encoding을 하는 등, 기본 전처리 과정을 거쳐, 선택된 모델들 학습한다. 이후, 학습된 모델 간의 교차 평가(Cross Validation)을 통해, 성능이 가장 좋은 모델을 선택하고, 해당 모델의 Test Dataset에 대한 예측 결과를 도출한다. EDA, 학습 및 평가 과정에 대한 자세한 내용은 2. Methods와 3. Result에서 살펴볼 수 있다.

	Attribute	Content
은행 고객 정보	**Age**	Numeric
	Job	Type of job (Categorical: 12 jobs)
	Marital	marital status(Categorical :4 types)
	Education	Categorical: 9 types
	Default	has credit in default? (3 types : 0, 1, unknown)
	Housing	has housing loan? (3 types : 0, 1, unknown)
	Loan	has personal loan? (3types, : 0, 1, unknown)
이전 분기 마케팅 활동 정보	**contact**	contact communication type (2 types : telephone, cell phone)
	Month	last contact month of year
	Day_of_week	last contact day of the week
	Duration	last contact duration
	Campaign	number of contacts performed during this campaign and for this client
	Pdays	number of days that passed by after the client was last contacted from a previous campaign
	Previous	number of contacts performed before this campaign and for this client
	Poutcome	outcome of the previous marketing campaign
사회 및 경제 현황 정보	**Emp.var.rate**	employment variation rate
	Cons.price.index	consumer price index
	Cons.conf.index	consumer confidence index
	Euribor3m	euribor 3 month rate
	Nr.employed	number of employees
예측 변수	**y**	has the client subscribed a term deposit? (binary : yes or no)

[Table 1. Attributes and its contents]

2. Methods

변수 구성 및 분포 확인 → 가설 설립 → EDA을 통한 인사이트 도출(시각화), 가설 통계 검증→ 전처리 → 모델 학습 → 모델 성능 확인 → 최종 모델 선정→ 예측 → 분류 결과를 통한 마케팅 전략 설립

이전 제출한 Proposal에서 제안한 전체 workflow이다. Python 코드를 통해서, 작업을 수행하였다.

2.1 EDA Hypothesis

먼저 전체 모든 변수의 분포를 확인하여, 이상여부를 관찰하였고, 사전 제안한 변수들과 관련된 가설들을 검증하였다. 사전 Proposal을 통해서 제안한 가설들은 다음과 같다.

1. 은행 고객 정보
   - "신용 상태를 가늠할 수 있는 변수들(job, default, loan, housing)에 대해 타겟 변수 간의 빈도수 차이가 있을 것이다."
2. 사회 및 경제 현황 정보
   - "정기 예금 예치한 사람들("y" = yes)의 Euribor3m 값이 더 클 것이다."
   - "정기 예금 예치한 사람들("y" = yes)의 cons.price.index 값이 더 작을 것이다."
   - "Euribor3m 값이 클수록 cons.conf.index 값이 더 작을 것이다."
   - "정기 예금 예치한 사람들("y" = yes)의 emp.var.rate 값이 더 클 것이다."
3. 이전 분기 마케팅 활동 정보
   - "정기 예금 예치한 사람들("y" = yes)의 duration이 길 것이다."
   - "정기 예금 예치한 사람들("y" = yes)의 previous값이 더 클 것이다."
   - "정기 예금 예치한 사람들("y" = yes)의 campaign값이 더 클 것이다."
   - "정기 예금 예치한 사람들("y" = yes)의 contact = cell phone의 빈도값이 더 클 것이다."

수치형 변수에 대해서는 종속 변수의 라벨 별로 평균 값을 barplot으로 시각화하여 비교하였다. 이후, 큰 차이를 보이는 변수에 대해서, 변수 간의 독립성을 가정한 뒤, t-test(Student's t test)를 통해 종속변수의 라벨 간의 분포를 비교하고자 했다. 먼저 t-test의 기본 가정인 정규성을 검증하였고, 정규성을 갖지 않는 경우, 비모수 검정(Non-parametric Test)을 통해서 분포 차이의 유의성을 검증하였다.
범주형 변수에 대해서는 종속 변수 별, 해당 범주형 변수 값에 대한 가능도(Likelihood probability)을 비교하였다. 특정 변수의 특정 값에 대한 라벨 별 빈도수가 큰 차이를 보인다면, 종속 변수에 대한 분류기로서의 성능이 좋을 것으로 가정하고, 이를 관찰하였다. EDA 결과는 3.1 EDA result 에서 살펴볼 수 있다.

2.2 Preprocessing

EDA를 통해서 예측 모델 형성을 위해서 중요변수들을 선택하고, 불필요한 변수들을 제거하였다. 선택한 변수와 제거한 변수들은 다음과 같다.[Table 2] 이후, 범주형 변수에 대해서 One hot Encoding을 통해 각 text로 이루어진 값들을 0,1로 바꾸어 주었다. 수치형 변수에 대해서는 변수 간의 다른 범주에 의한 영향력을 조정해주기 위해서 Scaling을 하였는데, 대표 두가지 방법 Min-Max-Scaling과 Standard Scaling을 모두 시도하였다. 각 Scaling 수식을 다음과 같다.

$$(1)\ Min-\max Scaling\ = \frac{X - Min}{Max - MIn}$$

$$(2)\ Standard\ Scaling = \frac{X - mean}{Standard\ Deviation}$$

시도한 두 가지 Scaling 방법 별로, 모델링을 시도하였으며, 성능이 좋은 Scaling 방식을 최종적으로 선택하였다. 선택한 변수 중 Marital과 Housing의 경우, 결측값을 의미하는 "Unknown"값 들이 있었는데, 해당

값의 전체 중 비율이 극히 작은 것을 관찰했다.(Marital의 경우, Unknown의 비율이 0.194%이었고, Housing의 경우, 2.4%이었다.) 해당 값들을 가진 데이터의 경우, 결측값으로 판단하여, 해당 행들을 삭제하여 처리하였다.

[Table 2. Variables which are selected or removed]

Selected Variable	Age, Job, Marital, Education, Housing, Contact, Month, Duration, Campaign, Previous, Poutcome, Emp.var.rate, Cons.conf.index,
Removed Varables	Days_of_week, loan, default, pdays, cons.price.index, euribor3m, nr.employed

제거한 변수에 대한 기준들은 다음과 같다.

- **Default** : unknown의 비율이 20%로 높아, 의미 있는 변수는 아니다. 더군다나 no, yes 대부분 default = yes로 몰려 있어, 분류를 위한 변수로 영향력이 없다.[**Figure 1,2 참고]**
- **Loan** : 대부분의 대답이 no인데, 각 라벨 별, loan = no인 데이터의 비율이 차이가 크지 않아, 분류를 위한 변수로 영향력이 없다. **[Figure3,4 참고]**
- **Days_of_week** : 월~금요일에 해당하는 각 라벨별 빈도수 차이가 없기에 분류를 위한 변수로 영향력이 없다. **[Figure 5 참고]**
- **Pdays** : 이전 홍보 활동 때 연락을 하지 않았을 경우 999로 처리되었는데, 그 비율이 다른 값들에 비해 너무 크고, 모두 라벨이 no이므로, 값이 클수록 no로 처리될 편향성이 있다. 따라서 해당 변수를 지운다. **[Figure 6]**
- **Cons.price.index , euribor3m, nr.employed** : euribor3m, emp.var.rate, cons.price.idx, nr.employed간 correlation이 높은 것을 알 수 있다. 다중 공선성을 제거하기 위해서 해당 변수들을 제거하고, 모델링에는 나머지 변수와 correlation이 높은 emp.var.rate만 사용하고, emp.var.rate와 상관계수가 높지 않은 "cons.conf.index"도 그대로 사용하였다.**[Figure 7 참고]**

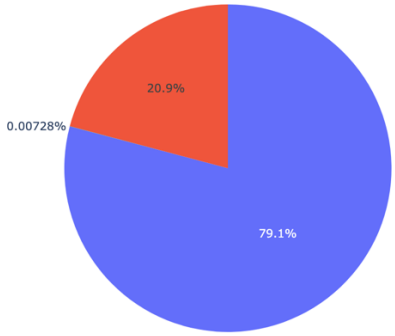
다른 변수들에 대해서도 추가적인 전처리 과정을 거쳤는데, Job의 경우, 총 12가지의 변수 값으로 구성되어 있었고, Education 변수 값에 대해서는 9가지 변수 값으로 구성되어 있었다. 이를 모두 One hot Encoding으로 변환 시에, 모델에 들어가는 변수의 수가 급격하게 증가하여 연산량이 증가하는 부작용이 있을 것을 우려하여, 변수 값들을 특정 그룹들로 묶어서 표현하였다.

- **Job (12 types)** : 경제적으로 안정한 직업과 안정하지 않은 직업으로 나누었다.
  - Stable : "admin.", "blue-collar", "technician", "services", "management", "entrepreneur"
  - Unstable : "retired", "self-employed", "housemaid", "unemployed", "student"
  - Unknown

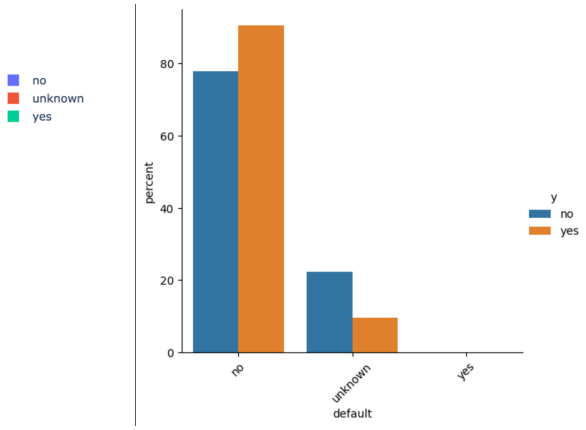
- **Education (9 types)**
  - Tertiary : university.degree, professional.course,high.school
  - Secondary : basic 4y, 6y, 9y
  - Primary : illiterate
  - Unknown

Month 변수에 대해서도 전처리 과정을 거쳤는데, 12개의 달 모두 One hot Encoding으로 나타내기보다는 다른 변수와의 관계를 파악하여 그룹화하였다. 특히 달 정보는 경제 지수(Euribor3m, emp.var.rate)와 관련이 있었는데, 경제 지표가 좋은 달이 있고, 안 좋은 달이 있던 것으로 파악하였다. 3, 4, 9, 10, 12월에는 특히 Euribor3m 지수가 작고, emp.var.rate가 작은 것을 확인하였다. **[Figure 8, 9 참고]**

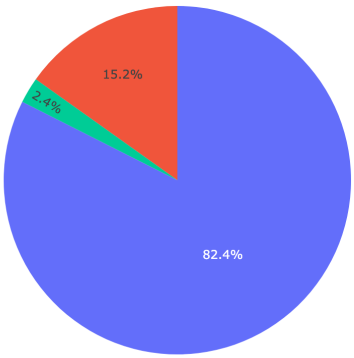
- **Month(12 types)**
  - Month_good : apr, oct, sep, mar, dec
  - Month_bad : may, jul, aug, jun, nov



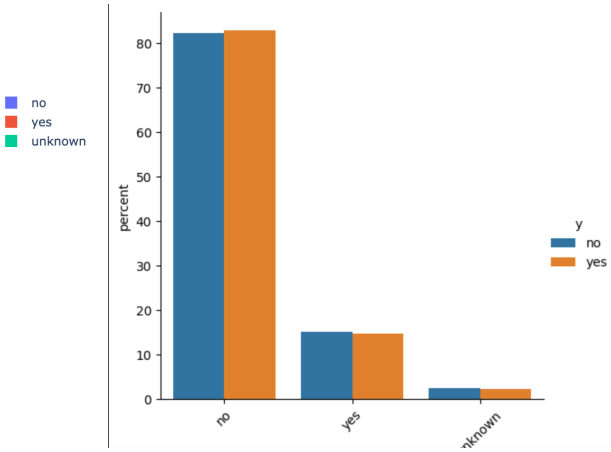
[Figure1. Pie chart of "Default"]



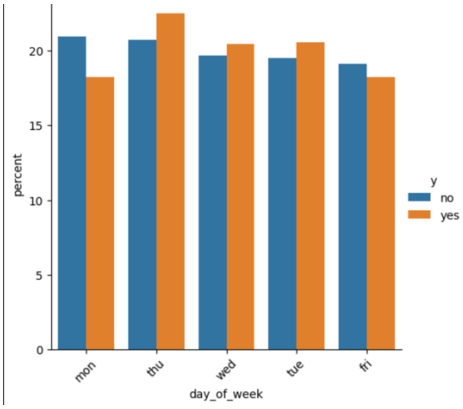
[Figure 2 Likelihood Probability comparison]



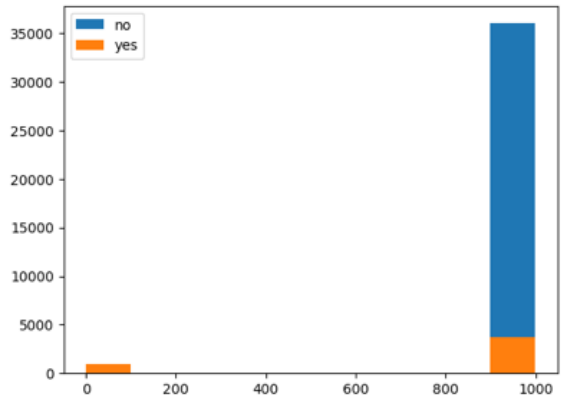
[Figure 3 Pie chart of "loan"]



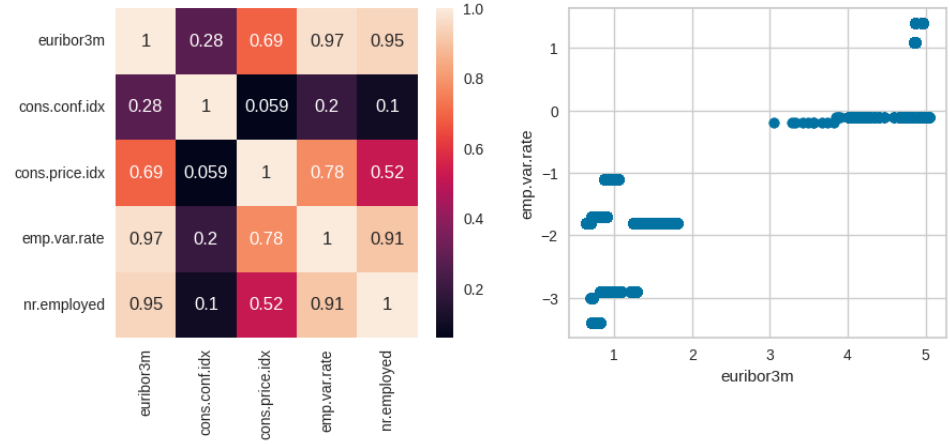
[Figure 4 Likelihood Probability comparison]



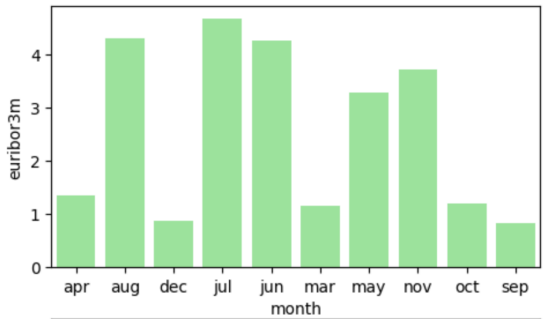
[Figure 5. Likelihood Probability of "Days of week"]



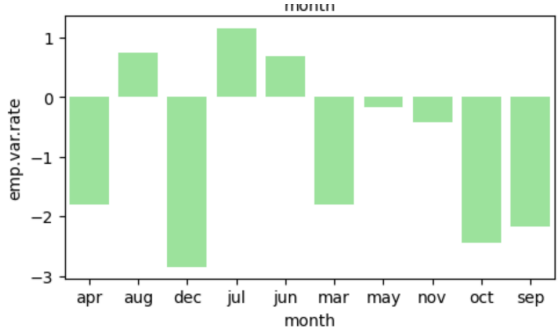
[Figure 6. Distribution of "pdays"]



[Figure 7. Correlation Heatmap between 4 variables (euribor3m, emp.var.rate, cons.price.idx, nr.employed)]



[Figure 8. Euribor3m monthly mean value]



[Figure 9. Emp.var.rate monthly mean value]

2.3 Training Set Ups

3.2에서 제시한 전처리를 거친 데이터 셋 총 32095 개의 데이터들을 이용하였다. 모델을 학습하기 전, 전체의 20%를 Test set으로 설정하여 별도로 떼어놓았다. 나머지 전체 80%를 학습 데이터로 사용하였고, 교차 평가(cross validation) 기법을 사용하여, 모델을 학습 및 성능 평가를 하였다 Fold 개수 5개로, 4개는 Train set으로, 1개는 Validation set으로 설정하여 총 5번의 학습 및 평가가 이루어지게 하였다. 특히 Fold를 형성 할 때는 Stratified Fold 방식을 채택하여, 기존 종속 변수(y)의 개수 차이에 의해서, 특정 라벨이 Train set 혹은 Validation set으로 편향되지 않도록 하였다.

2.4 Modeling

모델 학습에서 사용한 기본 모델은 총 4가지를 사용하였다. Logistic Regression, Decision Tree, SVM(Support Vector Machine), ANN(Artificial Neural Network)를 2.3 Training Set Ups에서 제시한 방법으로 모델 학습 이후, 평가 지표를 도출하여, 비교하였다. 각 모델의 성능은 3.2 Validation Result에서 확인할 수 있다. 기본 4가지 모델 말고, Ensemble 기법을 사용한 모델들도 학습시켜 성능을 확인하였다. Boosting 기법을 사용한 Catboost, LightGBM(Light Gradient Boosting Machine), GBC(Gradient Boosting Classifier), XGBoost(Extreme Gradient Boosting)을 사용해보았고, Feature selection에서 Bagging Ensemble 기법을 사용한 Random Forest Classifier을 사용하였다. 이후, 학습한 모델 중, validation 성능이 가장 좋은 5가지 모델들을 Stacking 기법과 Soft Voting 기법을 사용하여 최종 모델 성능 지표를 도출하였다.

** Stacking 기법 : 기본 모델들의 예측 결과값을 입력 받아, 최종 예측을 하는 메타 모델 형성 기법

** Soft Voting 기법 : 기본 모델들의 예측 결과값 중, Majority(다수결)을 결과로 도출하는 Hard Voting 과 달리, 예측 확률 값들을 평균 내어, 분류 예측 결과를 도출하는 기법이다.

2.5 Performance Metrics

기본 모델 성능 평가 기준은 정확도(Accuracy)로 정했으며, 비슷한 정확도 모델들 간의 비교는 F1 score이 더 높은 모델을 선택하였다. 또한 ROC(Receiver Operating Characteristic) curve을 그려보고, AUC(Area Under Curve) 값을 계산하여, 모델을 평가해보았다.

3. Results & Analysis

3.1 EDA result

2.1 EDA Hypothesis에서 제시한 가설들을 중점적으로 EDA을 통한 가설검증한 결과를 변수 종류 별로 살펴보면 다음과 같다.

3.1.1 은행 고객 정보

유일한 수치형 변수인 Age에 대해서 히스토그램[figure 8]을 그려보면 왜도(0.7),첨도(0.7)로 일반적인 왜도, 첨도에 의한 정규성 검증(왜도 3이하, 첨도 8이하)에 대해서는 정규성을 갖는 것으로 판단되었다. 하지만, Anderson-Darling test와 KS-Test에 의해 정규성 검정에 대한 p value값이 모두 0이 나온 것으로 미루어 보아, "정규성 분포이다."라는 귀무 가설을 기각하게 되어, 정규성은 띈다는 것은 유의하지 않은 것으로 판단된다. Age 변수의 각 종속변수 라벨 별 평균값을 비교하였을 때, 데이터의 분포가 특정 라벨로 편향되어 있지 않았다는 것도 알 수 있었다.**[Figure 10, 11]**

2.2 preprocessing 과정에서 제거한 변수들(default, loan)을 제외하고, 다른 변수에 대해, 예금 등록하는 사람들(y= yes) 과 예금을 등록하지 않은 사람들(y= no)의 비율을 비교한 결과 다음과 같이 결론을 내릴 수 있었다.

예금 등록을 하는 사람들(y= yes)에 대해서

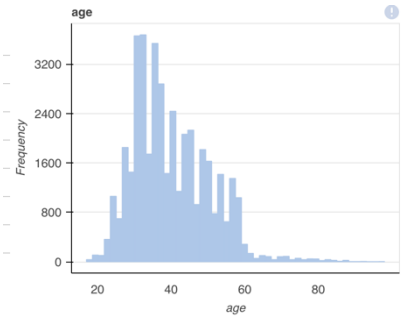
Marital → Marital 변수 값이 married보다 single인 고객들의 비율이 더 크다

Housing →집 대출이 있는 고객들이 없는 고객들보다 비율이 더 크다.

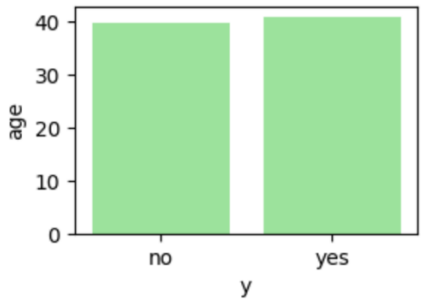
Job → stable한 직업을 가진 쪽의 비율이 더 낮았다.

Education → 교육을 더 받은 쪽의 비율이 더 높았다.

.



[Figure10 Histogram of Age]



[Figure11 Mean comparison of Age]

3.1.2 사회 및 경제 현황 정보

기존 수치형 변수에 대해 "Age" 변수 이외에 모두 정규성을 띄지 않은 것을 정규성 검정을 통해서 확인하였다. 따라서 정규성을 가정해야 하는 t-test보다 비모수 검정을 통해서 수치형 변수간의 차이가 있음을 확인하였다. 각 라벨별로 데이터를 구분하고, 이에 대해 비모수 검정을 한 결과 다음과 같았다. **[Figure 12]**

```
euribor3m
test for difference :  0.0
test result 'u1 : no > yes' : 0.0

cons.conf.idx
test for difference :  5.901951166896614e-17
test reuslt 'u1 : yes > no' : 2.950975583448307e-17

cons.price.idx
test for difference :  9.572608866919173e-136
test result 'u1 : no > yes' : 4.7863044334595866e-136

emp.var.rate
test for difference :  0.0
test result 'u1 : no > yes' : 0.0

nr.employed
test for difference :  0.0
test result 'u1 : no > yes' : 0.0
```

[Figure 12 Code results of Non-parametric tests]

- Euribor3m → y= no 인 사람들의 분포가 더 큰 값들을 가진다.
- Cons.Conf.index → y = yes 인 사람들의 분포가 더 큰 값을 가진다.
- Cons.Price.index → y = no 인 사람들의 분포가 더 큰 값을 가진다.
- Emp.var.rate → y=no 인 사람들의 분포가 더 큰 값을 가진다.
- Nr.employed → y= no 인 사람들의 분포가 더 큰 값을 가진다.

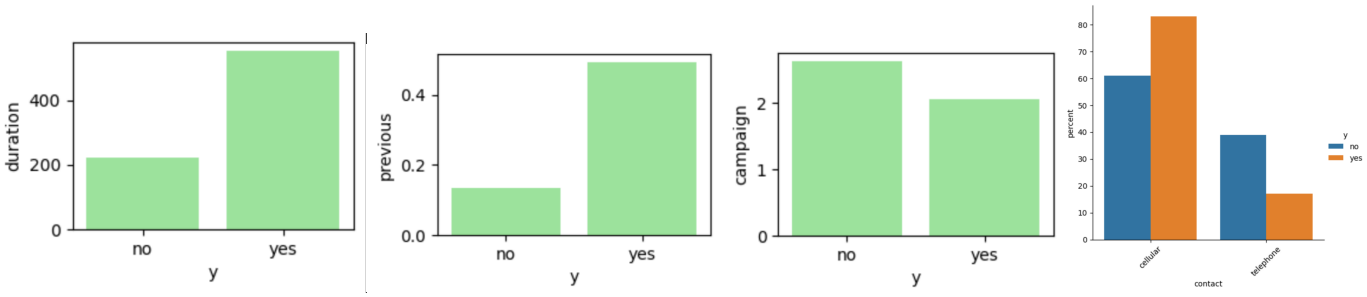
2.2 Preprocessing 과정을 거쳐서 다른 변수들을 제거한 후, 사용한 emp.var.rate 값의 결과를 살펴본 결과, emp.var.rate 값이 클 때, 즉 채용율이 작을 때, 예금 예치를 할 가능성이 더 적다는 가설을 검증할 수 있었다.

3.1.3 이전 분기 마케팅 활동 정보

2.1 EDA Hypothesis 에서 제시한 이전 분기 마케팅 활동 관련 변수에 대한 가설들은 다음과 같았다.

- "정기 예금 예치한 사람들("y" = yes)의 duration이 길 것이다."
- "정기 예금 예치한 사람들("y" = yes)의 previous값이 더 클 것이다."
- "정기 예금 예치한 사람들("y" = yes)의 campaign값이 더 클 것이다."
- "정기 예금 예치한 사람들("y" = yes)의 contact = cell phone의 빈도값이 더 클 것이다."

이에 대해, **[Figure 13]**을 통해서 모두 사실이었음을 확인할 수 있었다.



[Figure 13 Duration, Previous, Campaign, Contact]

3.2 Validation Result

2.4 Modeling에서 제시한 모델들에 대해 Fold 5개에 대한, cross validation을 통해 모델 학습 후, validation set에 대한 평가 지표들을 [Table 3]를 통해서 볼 수 있다. Catboost, LightGBM, GBC, XGBoost, Random Forest Classifier가 9개의 모델 상위 5개의 모델이며, 이를 stacking과 soft-voting 기법을 통해 ensemble하여 최종 예측 성능 또한 볼 수 있다.

[Table3]에서 9개 모델 중, 각 성능 지표 별 가장 높은 값을 볼드체로 표시했다. Min-Max Scaling 한 모델 중에서 가장 성능이 좋은 모델은 Catboost 이었으며, Standard Scaling 한 모델 중 가장 성능이 좋은 모델 역시 Catboost이었다. Min-Max Scaling과 Standard Scaling 중, 전처리 방법에 의한 성능을 비교한 결과, Decision Tree와.ANN, XGboost. 모델을 제외한 다른 6개 모델이 Standard Scaling을 한 경우 성능이 더 좋았던 것을 알 수 있다. Soft-Voting과 Stacking 간의 비교를 해보았을 때는, Soft-Voting한 모델이 더 성능이 더 좋았음 역시 알 수 있었다. 9개의 모델보다는 상위 5개의 모델 간 Ensemble 기법을 적용한 모델이 좋았기에, 최종 모델로 전처리 Standard Scaling, 상위 5개 모델의 Soft Voting을 적용한 모델로 결정하였다.

[Table3. Validation Performance of Models]

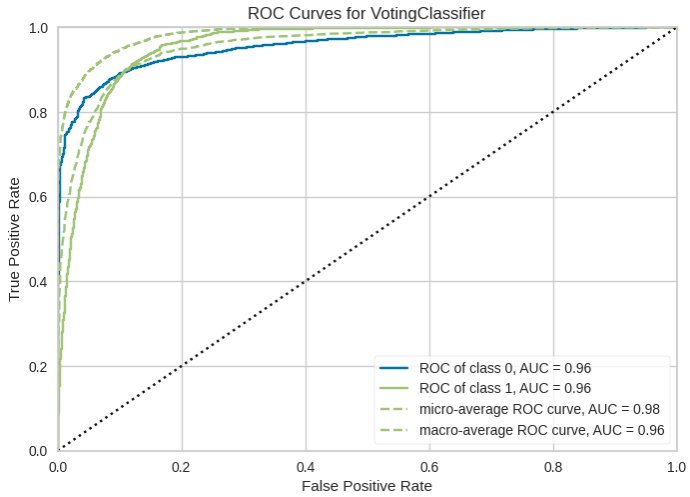
Model	Min-Max Scaling			Standard Scaling		
	Accuracy	F1 score	AUC	Accuracy	F1 score	AUC
Logistic Regression	0.9051	0.4466	0.9233	0.9055	0.4619	0.9246
Decision Tree	0.8831	0.4841	0.7113	0.8828	0.4778	0.7128
Support Vector Machine	0.9003	0.3683	0	0.9012	0.3809	0
ANN(Artificial Neural Network)	0.911	0.568	0.749	0.9107	0.548	0.750
Catboost	**0.9137**	0.5656	0.9420	**0.9145**	**0.5784**	0.9430
LightGBM	0.9125	**0.5700**	**0.9425**	0.9133	0.5612	**0.9436**
GBC	0.9124	0.5463	0.9415	0.9124	0.5463	0.9415
XGBoost	0.9103	0.5569	0.9375	0.9101	0.5565	0.9380
Random Forest Classifier	0.9088	0.5395	0.9310	0.9091	0.5401	0.9294
Soft-Voting Ensemble	**0.9143**	0.5670	**0.9438**	**0.9154**	**0.5862**	**0.9440**
Stacking Ensemble	0.9135	**0.5734**	0.9407	0.9151	0.5823	0.9430

3.3 Test Result

최종 모델로 선택된, Standard Scaling을 적용한 상위 5개 모델(Catboost, LightGBM, GBC, XGBoost, Random Forest Classifier)의 Soft-Voting Ensemble 모델을 이용하여, 기존 전체 데이터 셋의 20% 비율의 Test Set에 대한 예측값을 도출하였다. 이에 대한 최종 모델 성능은 [Table 4]와 [Figure 14]을 통해서 알 수 있다.

[Table 4 Test Performance of Sof Voting Ensemble Model]

	Top 5 Soft Voting Ensemble Model
Test sample number	8024
Accuracy	0.9126
Error Rate	0.08740
F1 score	0.5974
Confidence interval of accuracy	[0.906, 0.919]



[Figure14. ROC curve of Voting Ensemble Model]

4. Conclusion

은행 고객 정보, 사회 및 경제 현황 정보 및 이전 분기 마케팅 활동 정보 등이 담겨있는 데이터인 UCI의 Bank Marketing Campaign 데이터 셋을 통해 이번 분기 정기 예금 예치 여부를 예측하는 모델을 형성하였다. 20개의 변수 중, 불필요한 변수 7개를 제거한 후, 13개의 변수들을 그대로 사용하지 않고, 일련의 전처리 과정을 거쳤다. 이후, 범주형 변수에 대해서는 One hot Encoding을 수치형 변수에 대해서는 Min-Max Scaling과 Standard Scaling을 각각 적용하여 모델을 학습시켰다. 모델을 학습하기 전에, 중요한 변수들을 선별하기 위해서 EDA 과정을 거쳤는데, 사전에 각 변수에 대한 가설들을 설립한 후, 이를 검증하는 과정을 거쳤다.

다수의 가설은 사실임을 확인하였으나, 가설과 다른 부분도 확인 할 수 있었다. Euribor3m(3개월 간 유럽의 금리)가 가장 영향력이 있는 변수로 예상하였고, 금리가 높을 수록, 예금 예치 등록을 사람들이 비율이 더 클 것으로 예상하였다. 하지만 오히려 금리가 낮은 값일 때, 예금 예치 등록하는 사람들의 비율이 더 높았음을 확인하였다. 하지만, Emp.var.rate(employment variation rate) 값이 클 때, 예금 예치 등록하는 사람들의 비율이 더 낮았던 것을 미루어 보아, 채용률이 낮을 때, 경기 불황으로 인해, 예금을 예치하려는 사람이 적은 것은 예상과 다르지 않았다. 따라서 Euribor3m가 예상과 다르게 반대 경향을 보인 것에 대해, "금리 후반영", 즉 경기 불황일 경우, 금리를 높여 경기 안정을 취하거나, 경기가 과한 호황일 경우, 금리를 낮춰 경기를 안정 취하려는 정부의 조치에 의한 값으로 해석하였다.

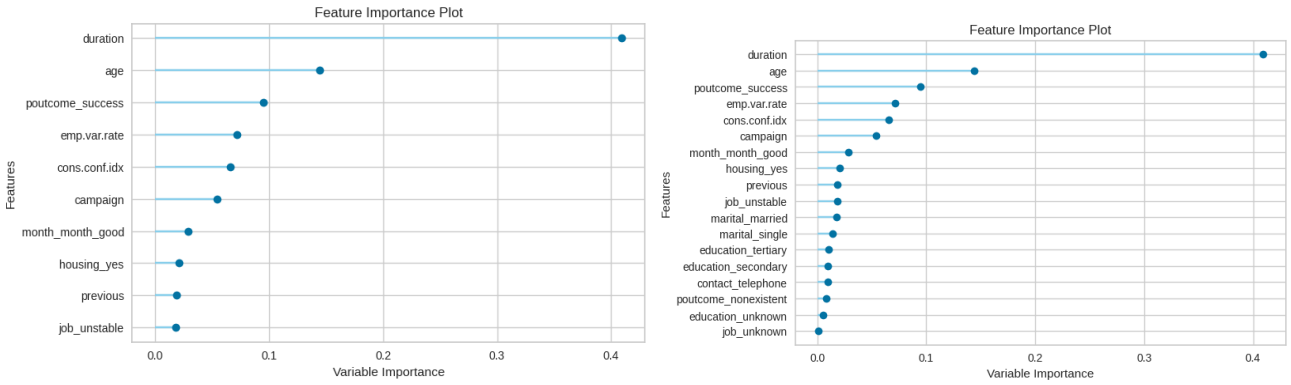
예측 모델 학습 결과, 보다 단순한 모델에 속하는 Logistic Regression, Decision Tree, SVM, ANN의 경우, 성능이 좋지 않았으나, 모델 자체에 Ensemble 기법을 적용한 모델이 더 성능이 좋았다. 이는 데이터 셋 변수들의 분포가 각 라벨별로 분명하게 구분되지 않았기 때문인 것으로 해석된다. 따라서 단순한 Decision Boundary 형성을 하는 단순 모델의 경우, 성능이 좋지 않았으며, 오히려 단순 모델들의 조합 및, 데이터 셋에 대한 임의 추출 방식을 채택하는 Boosting과 Bagging 기법을 사용하는 Catboost, LightGBM, GBC, XGBoost, Random Forest Classifier와 같은 모델들의 성능이 좋았다.

Tree 기반의 모델을 통해서는 각 변수들의 중요도를 파악할 수 있었는데, Decision Tree Classifier의 모델 학습 과정에서 각 변수들의 Importance를 관찰하였다. **[Figure 15]** 그 결과, Duration의 변수가 다른 변수와 차이가 분명하게 보였음을 알 수 있었다. 이를 통해서, 고객들의 예금 예치 등록을 유도하기 위해서, 다른 요인들보다 해당 홍보 활동의 지속성이 중요함을 알 수 있었다. 지속적으로 홍보 활동에 노출되는 고객일수록 다음 분기

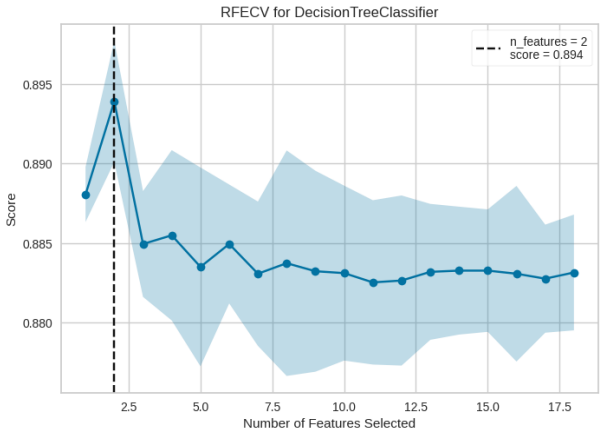
에 예금 예치를 할 확률이 높다.

해당 프로젝트를 통해서 기여한 바는 다음과 같다.

- EDA 과정을 통한 고객과 관련된 정보들 간의 관계를 파악하고, 예금 예치 여부와 관련된 경향들에 대한 가설들을 설립하여 이를 검증하였다.
- 금융 관련 데이터를 활용하여 예금 예치 등록 여부를 분류 예측하는 모델을 수립하였다.
- 예금 예치 여부에 영향을 미치는 주요 요인들을 파악하고, 이를 다음 홍보 활동 전략에 반영한다.



[Figure 15. Feature Importance plot from Decision Tree Classifier]



[Figure 16. RFECV for Decision Tree Classifier]

5. References

[1] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: Un iversity of California, School of Information and Computer Science.

[2] Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Ban k Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014