Data Mining for Business Analytics

Project(mid-term)

Predicting Software Reselling Profits

2023.04.17

Predicting Software Reselling Profits

소프트웨어 재판매 이익 예측

- 타이코 소프트웨어(Tayko Software)는 게임 및 교육용 소프트웨어를 판매하는 소프트웨어 카탈로그 회사이다. 이 회사는 소프트웨어 제품 제조로 창업하였고 나중에 제품에 대한 제3의 소유권을 가지게 되었다. 최근 이 회사는 새로운 카탈로그에 들어갈 제품 목록을 수정하였고, 이를 고객에게 우편 배송하였다.
- 이 우편물 발송으로 2,000건의 구매 성과를 올렸다. 이 데이터를 기반으로 구매 고객의 소비금액을 예측하는 모델을 고안하고자 한다.
- [Tayko.csv] 파일은 2,000건에 대한 구매 정보를 포함하고 있다. 아래의 표는 이 문제에서 사용된 변수들에 대하여 기술한 것이다. (엑셀 파일에는 추가적인 변수들이 포함되어 있음)

변수 이름	변수 내역
US	미국 주소인지에 대한 여부
Freq	전년도의 거래 건수
last_update_days_ago	고객레코드 최종갱신일로부터의 경과 일수
Web order	고객이 최소한 한 번 이상 인터넷 구매를 했는가에 대한 여부
Gender=male	남성(1) 또는 여성(0)
Address_is_res	거주지 주소인지에 대한 여부
Spending (결과 변수)	테스트 우편물에 의한 구매액(달러)

Predicting Software Reselling Profits

소프트웨어 재판매 이익 예측

a. 범주형 변수들에 대한 테이블을 만들고, 각 범주별로 소비금액의 평균과 표준편차를 계산하시오.

b. 연속형 변수들에 대하여 산점도(2개)를 작성하여 소비금액과의 관계를 탐색하시오(Spending 대 Freq, Spending 대 last_update_days_ago). 이들이 선형관계가 있어 보이는가?

c. Spending에 대한 예측모델을 적합시키기 위해:

   1) 2,000개의 레코드를 학습 데이터와 검증 데이터로 나누시오.

   2) Spending을 결과변수로 설정하고 위 표의 6개 예측변수를 사용하여 다중 선형회귀 모델을 만드시오. 추정된 회귀모델식을 구하시오.

   3) 이 모델을 기반으로 하였을 때, 가장 많은 돈을 지출할 것 같은 구매고객의 유형은 무엇인가?

   4) 예측변수들의 수를 줄이기 위하여 후진제거 방법을 사용한다면, 어떠한 예측변수가 모델로부터 가장 먼저 탈락되겠는가?

   5) 검증 데이터의 첫 번째 구매 데이터를 이용하여 예측값과 예측오차가 어떻게 계산되는지 보이시오.

   6) 검증 데이터에 대한 모델의 성능을 검토한 후, 모델의 예측 정확도에 대하여 평가하시오.

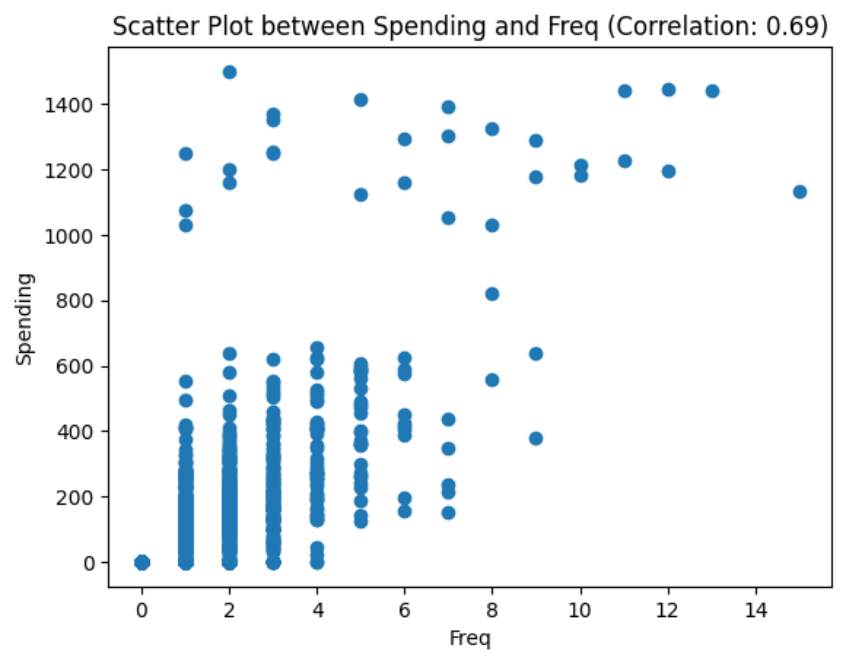
   7) 모델의 잔차에 대한 히스토그램을 작성하시오. 정규분포를 따르는가? 이는 모델의 예측 성능에 어떠한 영향을 미치는가?

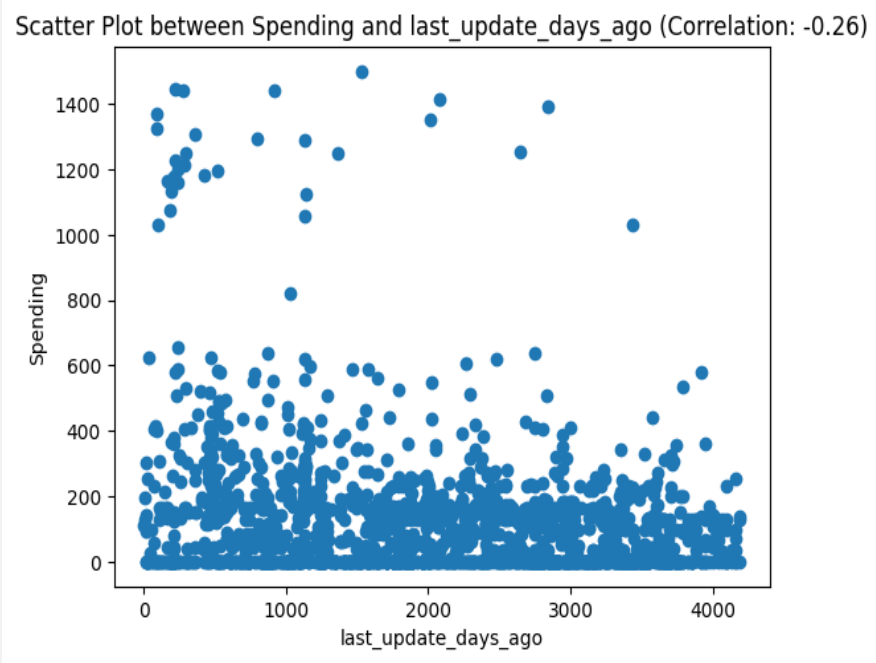
Predicting Software Reselling Profits

A.

Predicting Software Reselling Profits

B. 소비 금액 상관관계 - 상관분석





전년도 거래 건 수와 결과변수는 양의 상관관계로
거래 건 수가 적을 수록 구매액이 낮고
거래 건 수가 높을 수록 구매액이 높음

고객레코드 최종 갱신일로부터의 경과 일수와
결과변수는 음의 상관관계

Predicting Software Reselling Profits

C. Spending에 대한 예측모델

1)

```
# 데이터 분할
X = pd.get_dummies(tayco_df[predictors])
y = tayco_df[outcome]
train_X, valid_X, train_y, valid_y = train_test_split(X, y, test_size=0.4, random_state=1)
```

학습 데이터: 60% / 검증 데이터: 40%

2)

```
# 컬럼 이름 변경
tayco_df = tayco_df.rename(columns={'Web order': 'web_order', 'Gender=male': 'gender'})

predictors = ['US', 'Freq', 'last_update_days_ago', 'web_order', 'gender', 'Address_is_res']
outcome = 'Spending'

# 데이터 분할
X = pd.get_dummies(tayco_df[predictors])
y = tayco_df[outcome]
train_X, valid_X, train_y, valid_y = train_test_split(X, y, test_size=0.4, random_state=1)

#선형 회귀 모델 생성 및 학습
tayco_lm  = LinearRegression()
tayco_lm .fit(train_X, train_y)

# print coefficients
print('intercept ', tayco_lm .intercept_)
print(pd.DataFrame({'Predictor': X.columns, 'coefficient': tayco_lm .coef_}))

# 성능 측정 출력
regressionSummary(train_y, tayco_lm .predict(train_X))
```

```
intercept  10.17629741458822
             Predictor  coefficient
0                   US    -4.620293
1                 Freq    91.274450
2  last_update_days_ago    -0.010374
3            web_order    18.628731
4               gender    -9.111366
5        Address_is_res   -75.815354

Regression statistics

                Mean Error (ME) : 0.0000
Root Mean Squared Error (RMSE) : 125.9999
     Mean Absolute Error (MAE) : 79.4772
```

```
Spending = 10.18 + (-4.62 * US) + (91.27 * Freq) + (-0.01 * last_update_days_ago) + (18.63 * web_order) + (-9.11 * gender) + (-75.82 * Address_is_res)
```

Predicting Software Reselling Profits

C. Spending에 대한 예측모델

3) 전년도 거래 건 수가(Freq) 많은 구매 고객이 가장 많은 돈을 지출 할 것으로 예상됨

4) 후진제거 시 미국 주소지 여부(US)가 가장 먼저 탈락됨

```python
def train_model(variables):
    model = LinearRegression()
    model.fit(train_X[variables], train_y)
    return model

def score_model(model, variables):
    return AIC_score(train_y, model.predict(train_X[variables]), model)

best_model, best_variables = backward_elimination(train_X.columns, train_model, score_model, verbose=True)

print(best_variables)
```

```
Variables: US, Freq, last_update_days_ago, web_order, gender, Address_is_res
Start: score=15028.53
Step: score=15026.76, remove US
Step: score=15026.38, remove gender
Step: score=15026.38, remove None
['Freq', 'last_update_days_ago', 'web_order', 'Address_is_res']
```

Predicting Software Reselling Profits

C. Spending에 대한 예측모델

5)

	Predicted	Actual	Residual
674	89.214915	0	-89.214915

예측된 구매 금액: 674 / 실제 구매 금액: 89.21 / 잔차: -89.21

6)

```
Regression statistics

                Mean Error (ME) : 7.1933
Root Mean Squared Error (RMSE) : 136.7397
     Mean Absolute Error (MAE) : 83.6010
```

후진제거 전 모델 성능 평가 결과

```
Regression statistics

                Mean Error (ME) : 6.9616
Root Mean Squared Error (RMSE) : 136.5274
     Mean Absolute Error (MAE) : 83.4472
```

후진제거 후 모델 성능 평가 결과

- 후진제거 전 모델과 제거 후 모델 성능 평가 비교 결과 예측 오차의 평균은 7.19에서 6.96으로 감소
- RMSE와 MAE는 큰 변화는 없음

Predicting Software Reselling Profits

C. Spending에 대한 예측모델

7)