Data Mining for Business Analytics

Project(mid-term)

Predicting Software Reselling Profits

2023.04.17.

박영제(2022254001)

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. (식을써라)
6) 검증 데이터에 대한 모델의 성능을 검토한 후, 모델의 예측 정확도에 대하여 평가하시오. RMSE
7) 모델의 잔차에 대한 히스토그램을 작성하시오. 정규분포를 따르는가? 이는 모델의 예측 성능에 어떠한 영향을 미치는가?

Predicting Software Reselling Profits

범주형은 01만?

소프트웨어 재판매 이익 예측

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

테이블

US	source_a	source_c	source_b	source_d	source_e	source_m	source_o	source_h	...	source_x	source_w	Freq	last_update_days_ago	1st_update_days_ago	Web order	Gender=male	Address_is_res	Purchase	Spending
1	0	0	1	0	0	0	0	0	...	0	0	2	3662	3662	1	0	1	1	128
1	0	0	0	0	1	1	0	0	...	0	0	0	2900	2900	1	1	1	0	0
1	0	0	0	0	0	2	0	0	...	0	0	2	3883	3914	0	0	2	1	127
1	0	1	0	0	0	3	0	0	...	0	0	1	829	829	0	1	3	0	0
1	0	1	0	0	0	4	0	0	...	0	0	1	869	869	0	0	4	0	0
..	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1	0	0	0	0	0	1995	0	0	...	0	0	1	1701	1701	1	0	1995	0	1	30
1	0	0	0	0	0	1996	0	0	...	0	0	1	2633	2633	1	1	1996	0	1	10
1	0	0	0	0	0	1997	0	0	...	0	0	0	3394	3394	0	0	1997	0	0	0
1	0	0	0	0	0	1998	0	0	...	0	1	1	253	253	0	1	1998	1	0	0
1	0	0	0	0	0	1999	0	0	...	0	0	1	1261	1844	0	0	1999	0	0	0

[2000 rows x 25 columns]

표준편차 nan의 경우 한 개의 레코드만 있다. 이 경우 의미 없음으로 배제할것인지?

US	Spending mean	Spending std
0	101.216524	174.844401
1	102.924803	189.275664

Freq	mean	std
0	0.000000	0.000000
1	66.322476	104.424412
2	123.479714	151.509696
3	234.993243	226.259754
4	306.061224	165.153642
5	459.862069	270.661719
6	556.750000	344.195009
7	642.125000	520.126478
8	933.500000	324.925120
9	870.500000	433.828307
10	1199.000000	21.213203
11	1334.000000	151.320851
12	1320.500000	177.483802
13	1443.000000	NaN
15	1133.000000	NaN

last_update_days_ago	mean	std
1	109.000000	NaN
7	129.000000	NaN
9	196.000000	NaN
14	303.000000	NaN
15	71.000000	100.409163
...	...	...
4065	64.750000	58.987993
4096	75.666667	93.681731
4127	17.500000	24.748737
4157	75.666667	95.516840
4188	88.000000	76.374079

Web order	mean	std
0	82.902439	173.417088
1	129.199531	200.463840

Gender=male	mean	std
0	107.339642	190.83233
1	98.350810	183.02006

Address_is_res	mean	std
0	105.306162	199.521159
1	93.174208	132.204281

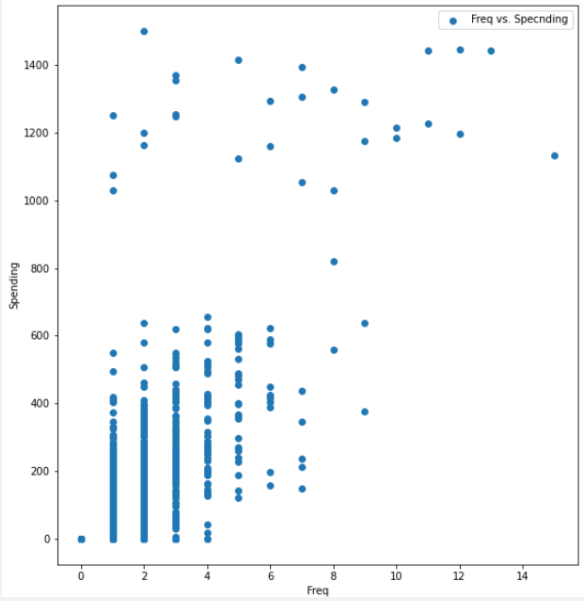
Predicting Software Reselling Profits

소프트웨어 재판매 이익 예측

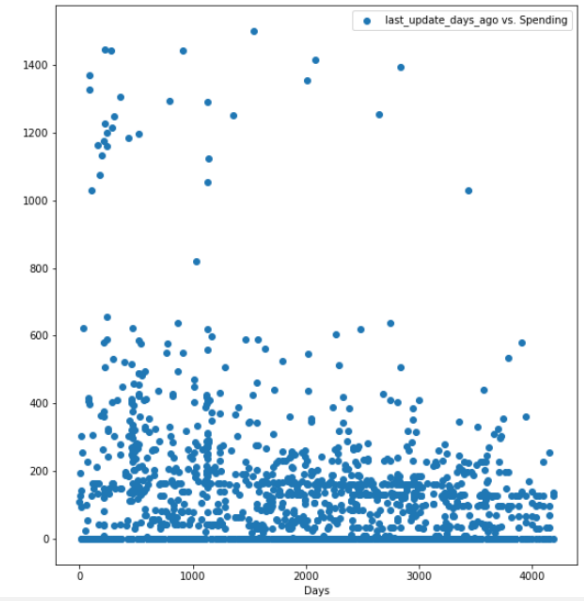
b. 연속형 변수들에 대하여 산점도(2개)를 작성하여 소비금액과의 관계를 탐색하시오(Spending 대 Freq, Spending 대 last_update_days_ago). 이들이 선형관계가 있어 보이는가?

```
fig, ax = plt.subplots(1, 2, figsize = (20, 10))
ax[0].scatter(soft_df['Freq'],soft_df['Spending'],label = 'Freq vs. Specnding')
ax[0].set_xlabel('Freq'), ax[0].set_ylabel('Spending')
ax[0].legend()
ax[1].scatter(soft_df['last_update_days_ago'],soft_df['Spending'],label = 'last_update_days_ago vs. Spending')
ax[1].set_xlabel('Days'), ax[0].set_ylabel('Spending')
ax[1].legend()
plt.show()
```

로그스케일을 적용해서 자세히 보자
Np.log()



VS Freq 선형관계가 있다고 판단됨.



VS Last update days ago 선형관계가 없다고 판단됨.

Predicting Software Reselling Profits

소프트웨어 재판매 이익 예측

C. Spending에 대한 예측모델을 적합시키기 위해:

1) 2,000개의 레코드를 학습 데이터와 검증 데이터로 나누시오.

```
X_train, X_valid, y_train, y_valid = train_test_split(X[:2000], y[:2000], test_size=0.2, random_state=42)
print("Train set size:", len(X_train))
print("Validation set size:", len(X_valid))
```

```
Train set size: 1600
Validation set size: 400
```

Predicting Software Reselling Profits

소프트웨어 재판매 이익 예측

C. Spending에 대한 예측모델을 적합시키기 위해:

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

```
df = pd.read_csv('Tayko.csv')
X = df[['US', 'Freq', 'last_update_days_ago', 'Web order', "Gender=male", 'Address_is_res']]
y = df['Spending']

model = LinearRegression()
model.fit(train_X, train_y)


print('intercept ', model.intercept_)
print(pd.DataFrame({'Predictor': X.columns, 'coefficient': soft_lm.coef_}))
regressionSummary(train_y, model.predict(train_X))
```

```
intercept  10.176297414608456
              Predictor  coefficient
0                    US    -4.620293
1                  Freq    91.274450
2  last_update_days_ago    -0.010374
3             Web order    18.628731
4           Gender=male    -9.111366
5        Address_is_res   -75.815354
```

다중선형회귀모델 수식 y = b0 + b1 * x1 + b2 * x2 + ... + bn *

Spending = 10.17 + (-4.62) * US + 91.27 * Freq + (- 0.01) * last_update_days_ago + 18.62 * Web order + (- 9.11) * Gender=male +(- 75.81) * Address_is_res / 실제 계산 해 볼것

Predicting Software Reselling Profits

소프트웨어 재판매 이익 예측

C. Spending에 대한 예측모델을 적합시키기 위해:

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

변수 이름	변수 내역	추정계수	긍정/부정
US	미국 주소인지에 대한 여부	-4.620293	부정
Freq	전년도의 거래 건수	91.274450	긍정
last_update_days_ago	고객레코드 최종갱신일로부터의 경과 일수	-0.010374	관계약함
Web order	고객이 최소한 한 번 이상 인터넷 구매를 했는가 여부	18.628731	긍정
Gender=male	남성(1) 또는 여성(0)	-9.111366	여성
Address_is_res	거주지 주소인지에 대한 여부	-75.815354	부정

우선순위 정열 시 다음과 같이 표현 가능
전년도 거래자, 거주지 주소가 아닌 자, 인터넷 구매이력이 있는 자, 여성, 미국 주소가 아닌 자

Predicting Software Reselling Profits

소프트웨어 재판매 이익 예측

C. Spending에 대한 예측모델을 적합시키기 위해:

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

의미가 없는 컬럼을 우선제거
어떤 컬럼이 의미가 없는가?

Predicting Software Reselling Profits

소프트웨어 재판매 이익 예측

C. Spending에 대한 예측모델을 적합시키기 위해:

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

숫자를 넣고 실제로 계산 해 봐라

Predicting Software Reselling Profits

소프트웨어 재판매 이익 예측

C. Spending에 대한 예측모델을 적합시키기 위해:

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

Predicting Software Reselling Profits

소프트웨어 재판매 이익 예측

C. Spending에 대한 예측모델을 적합시키기 위해:

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

회귀분석 자체에 문제가 있는지 여부 / 정규분포를 따르면 문제가 없는것
0에 많이 몰려 있을수록 예측성능이 좋다