Multiple linear regression

Features : 'n' number of features

label : 1 label → numeric

Venture Capitalist Use-case :

R&DSpend → numeric — f1

Administration → numeric — f2

Marketing → numeric — f3

State → categorical — f4

Profit → numeric — label

} No missing data issue

After handling categorical data,

California

Florida

New York

R&D Spend

Administration

Marketing

Profit

Equation of line goes:

Profit = b_0 + b_1 [California]

+ b_2 [Florida] + b_3 [NY]

+ b_4 [R&D Spend] + b_5 [Adm]

+ b_6 [Market]

FEATURE ENGINEERING

- Feature Selection
  - (a) RFE
  - (b) Select by model
  - (c) ANOVA (Analysis of Variance)
- Feature Extraction
  - (a) PCA (Principal Component Analysis)
  - (b) LDA (Linear Discriminant Analysis)

Feature Selection analysis

Regression :-

1. Correlation analysis (corr) ✓
2. Elimination using OLS ←
3. Feature selection ✓
4. Feature extraction

Backward elimination using OLS

Hypothesis testing

Significance level → 0.01 ← 0.05 ← 0.1 ←
(the amount of error you can tolerate)

p-value →

confidence → 0.99 ← 0.95 ← 0.9 ←
(the trust you can have on the model)

R&D ⟷ Profit

P value

Calc^{n} using OLS

$P <= 0.05$ → Rel^{n}ship exist so keep the feature

else → Eliminate feature

my significance level is

$p = 0.05$ ← Data Scientist

Role of ML engg:

- Eliminate the features whose p value is greater than 0.05

Ideal steps we perform in feature elimination using OLS / AIC

① Perform All-IN (because you provide all coeff and intercept coeff)

$$\text{profit} = (1)b_0 + b_1(\text{California}) + b_2(\text{Florida}) + b_3(\text{NY}) + b_4(\text{RED}) + b_5(\text{adm}) + b_6(\text{mark})$$

② Decide the Significance level (0.05)

③ Select the feature in summary which has highest p-value.

④ If pvalue > SL(0.05), eliminate that feature.

⑤ Repeat step 3 & 4 till condition ④ is false. If condition ④ is false, go to step 6.

⑥ Consider all remaining features as selected feature for model building.