**ANALYSING AND CREATING PREDICTIVE MODELS**

**USING LINEAR REGRESSION AND PCA**

**Importing Data and Exploring data**

To start exploring let’s first import the data into application using below Code:

SME\_data<-read.csv("SME\_Profit.csv") #load data

A screenshot of a computer

Description automatically generated

After importing we can see there 50 observations and 18 variables to get deeper understanding let’s explore the data:

str(SME\_data) #Structure of the dataA screenshot of a computer

Description automatically generated

we can see that data type of state is chr instead of factor. So lets change the data type using below code

SME\_data <- SME\_data %>%

mutate(State = as.factor(State))

summary(SME\_data) # Explore Data

A screenshot of a computer program

Description automatically generated

Summary of a dataset gives us descriptive analysis of the each column of the data set as the dataset is small instead of calculating individual column, I calculated all the columns at once.

Observations:

1. As we can see out of all 18 variables only 5 variables have the actual values and the rest is redundant data that i removed using the below code.

SME\_data<-SME\_data[,1:5] # Removing the redundant variables

str(SME\_data) #To View Structure of the data(Analyse the result)

A screenshot of a computer

Description automatically generated

Now we Can see there only 5 Variable rest all are deleted.

1. Every other variable except state (which is categorical Variable),as the outliers has the max value of the variable are greater than 3rd Quadrant value, but I am including them for the rest the procedure as I believe they are not that high compared their respective medians.

As we removed the redundant columns let’s analyze rest numerical variables using boxplot

Below to code is use to extract numerical variables:

numeric\_SME\_data<-Filter(is.numeric,SME\_data) #Creating dataframe for boxplots extracting only numeric variables

#Normalization

numeric\_SME\_data[, c("R.D.Spend", "Administration", "Marketing.Spend","Profit")] <-

scale(numeric\_SME\_data[, c("R.D.Spend", "Administration", "Marketing.Spend","Profit")])

#Boxplots

boxplot(numeric\_SME\_data,main = "Box Plots of Numeric Variables", col = c("lightblue", "lightgreen", "lightpink"))

A diagram of a box plot

Description automatically generated

We can observe there are no extreme outliers in the all the variables. Data does not any missing values we can start looking at relationship between the variables.

**Relationship Analysis**:

Let’s prepare scatter diagram of the all the variables in relation to Profit for understanding the relationship between them.

#Scatter Diagram

# Define a color palette for the states

state\_colors <- rainbow(length(levels(SME\_data$State)))

#partioning the plotting space

par(mfrow = c(1,3))

plot(SME\_data$Profit,SME\_data$R.D.Spend,col = state\_colors,

xlab = "Profit", ylab = "R.D.Spend")

plot(SME\_data$Profit,SME\_data$Administration,col = state\_colors, main = "Relationship Scatter Diagram",

xlab = "Profit", ylab = "Administration")

# Add legend

legend("topright", legend = levels(SME\_data$State),

col = state\_colors,

pch = 1)

plot(SME\_data$Profit,SME\_data$Marketing.Spend,col = state\_colors,

xlab = "Profit", ylab = "Marketing.Spend")

A white sheet with black text and colorful dots

Description automatically generated with medium confidence

Let's analyze the scatterplots based on the visual patterns in each graph:

**1. R&D Spend vs. Profit (Left Plot):**

* The points in this scatterplot form a clear upward trend, where higher R&D spend is associated with higher profit.
* The points seem to follow a roughly straight line, indicating a **strong positive linear relationship** between R&D Spend and Profit.
* Thus, this plot suggests a **linear relationship**.

**2. Administration vs. Profit (Middle Plot):**

* The scatterplot here does not show a clear trend. The points are scattered randomly with no obvious upward or downward pattern.
* This suggests that there is **no clear linear relationship** between Administration expenses and Profit. The relationship may be weak or even non-existent.

**3. Marketing Spend vs. Profit (Right Plot):**

* There is a slight upward trend, but the points are more spread out compared to the R&D Spend plot.
* While there might be some positive relationship, it does not seem as strong or linear as the one between R&D Spend and Profit. There is more variability in the points, indicating the relationship may be **weakly linear** or possibly influenced by other factors.

**Summary:**

* **R&D Spend vs. Profit**: Strong linear relationship (positive).
* **Administration vs. Profit**: No clear linear relationship.
* **Marketing Spend vs. Profit**: Possibly a weak linear relationship, but not as strong as R&D Spend.

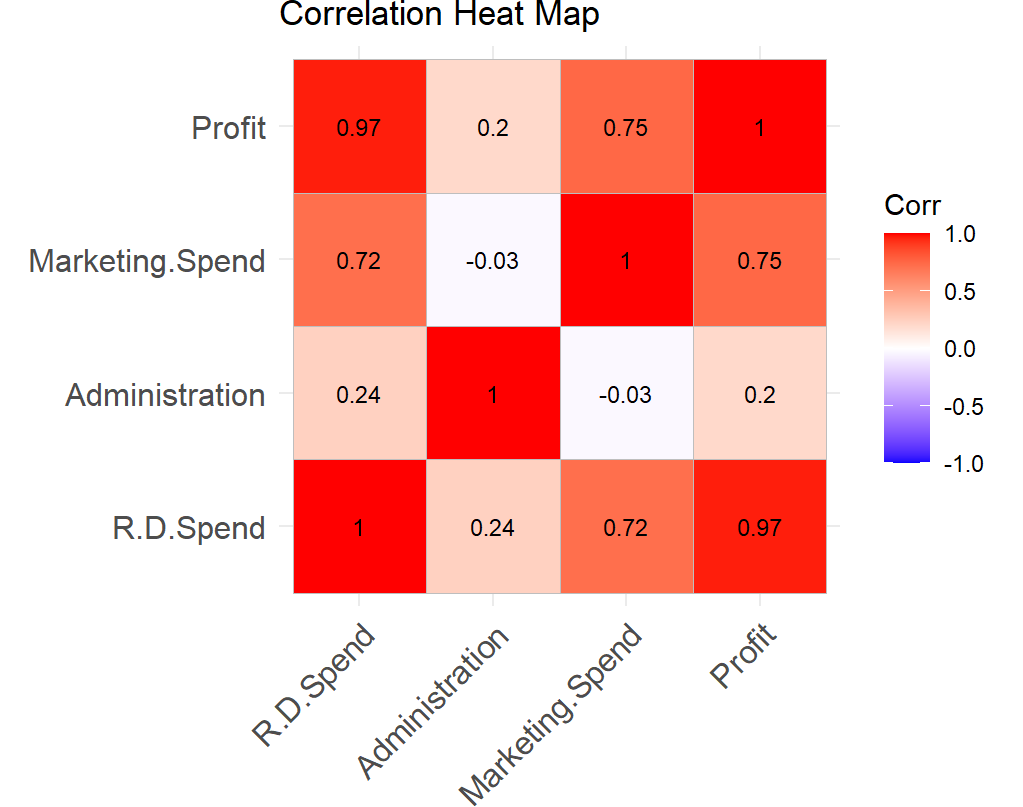
Correlation helps us to understand the relationship between variables and strong is the relationship in between the variables. The code creates a heatmap of correlations between different variables

install.packages("ggcorrplot")

library(ggcorrplot)

# Create a heatmap with percentage labels

ggcorrplot(cor(numeric\_SME\_data),title = "Correlation Heat Map", lab = TRUE, lab\_size = 3)



Observations:

1. R.D.Spend and Profit have a very strong positive correlation of 0.97(Almost perfect).
2. R.D Spend and Marketing.spend have a strong positive correlation of 0.72.
3. Profit and Marketing.spend have a strong positive correlation of 0.75.

**Creating Dummies Variables:**

As State is categorical variable let’s create dummy variables for the state. Since the if the state is not New York or California means it is Florida I am not adding separate column for Florida.

SME\_data$State\_Newyork<-ifelse(SME\_data$State=="New York",1,0)

SME\_data$State\_California<-ifelse(SME\_data$State=="California",1,0)

SME\_data<-SME\_data[,-4]

A screenshot of a computer

Description automatically generated

**Fitting a Predictive Model:**

Below code is used to create a predictive model for the entire dataset (i.e., SME\_data)

Profit\_model<-lm(Profit~.,SME\_data)

summary(Profit\_model) A screenshot of a computer screen

Description automatically generated

**Partitioning the Data:**

The following code is used to partition the data into training data and Validation which used further in creating the Model.

set.seed(1) # to guarantee that the same random values are produced each time you run the code.

Train.index<-sample(c(1:50),30) # used 60% of data for training(50\*60%)

train\_SME<-SME\_data[Train.index,] #training dataset

valid\_SME<-SME\_data[-Train.index,] #validation Dataset

A screenshot of a computer

Description automatically generated

**Multiple Linear Regression:**

As we have training and validation datasets. Let’s create the regression model using the below syntax.

SME\_data.lm<-lm(Profit~.,data=train\_SME)

options(scipen = 999)

summary(SME\_data.lm)

A screenshot of a computer

Description automatically generated  
**Predicting the profit using validation data:**

We already built the regression model so the next step for this is predicting the profit values for validation data using regression model of training data. Below code is used predict values for the profit values

**A screen shot of a computer

Description automatically generated**

**Residual Plot:**

Residual plot helps to know model fit and weather to check the underlying assumptions are right. The below code is used to draw a residual plot.

A graph with blue dots and a red line

Description automatically generated

**MLR and residual Comments:**

**1. Residuals:**

* The residuals show the differences between the actual and predicted values of Profit.
* The minimum residual is -30,448.9, and the maximum is 10,927.3. While the residuals are not perfectly centered around zero, the median residual of 445.6 suggests the model isn't far off on average.
* There seems to be a large spread between the minimum and maximum residuals, which may indicate potential outliers or heteroscedasticity (non-constant variance of residuals).
* The residual plot looks **reasonable** for a linear regression model. The residuals are randomly dispersed without a clear pattern, suggesting that the assumptions of linearity and constant variance (homoscedasticity) are met.

**2. Coefficients:**

* **Intercept**: The intercept is significant with a p-value of 0.00003, meaning when all predictor variables are zero, the model predicts a baseline profit of 48,567.67.
* **R&D Spend**: This variable has the highest t-value (15.724) and is highly significant (p-value close to zero). A unit increase in R&D Spend is associated with an increase in profit by 0.82557 units, making it the strongest predictor of Profit in the model.
* **Administration**: The coefficient for Administration is negative (-0.06955), suggesting a slight inverse relationship with profit, but it is not statistically significant (p-value = 0.329). This variable may not contribute much to predicting profit.
* **Marketing Spend**: The positive coefficient (0.03115) indicates a small positive effect on profit, but with a p-value of 0.137, it is not statistically significant.
* **State Variables (New York, California)**: Neither of the state dummy variables is significant, with p-values of 0.199 (New York) and 0.467 (California), implying that the state location does not have a strong impact on profit in this model.

**3. Significance Codes:**

* The \*\*\* symbols indicate the level of significance. Only the intercept and R&D Spend are statistically significant at the 0.001 level.

**4. Model Fit:**

* **Residual standard error** of 9263 indicates the average difference between the observed and predicted profits.
* **Multiple R-squared** of 0.9666 suggests that 96.66% of the variance in Profit is explained by the model, which indicates a very good fit.
* **Adjusted R-squared** of 0.9596 adjusts for the number of predictors and still suggests a high level of explained variance.
* The **F-statistic** (138.9, p-value extremely close to 0) indicates that the overall model is statistically significant.

**Overall Conclusion:**

The model shows a strong fit with R&D Spend being the most significant predictor of Profit, while Administration, Marketing Spend, and state variables do not significantly contribute to the model.

**Selecting subsets of Predictors:**

As we already create a Multiple linear regression model with all the predictors, but its not true that using all predictors to create a model guarantee the best results. Hence let’s use subsets of variable to create models using different approaches and find the best fitted model after the carefully their respective performances.

1. **Exhaustive R:**

This method assess all possible subsets of predictors. (single, pairs,

triplets, etc.) and the search gives us the best subsets of predictors.

###Exhaustive R###

Exhaustive\_search <- regsubsets(Profit~.,data = train\_SME,

nbest = 1,nvmax = dim(train\_SME)[2],method = "exhaustive")

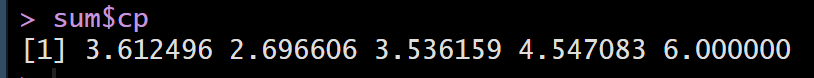
sum<-summary(Exhaustive\_search)

sum

A computer screen with white text

Description automatically generated

sum$cp



1. **Forward R:**

Code:

## create model with no predictors for bottom of search range

SME\_data.lm.null <- lm(Profit~1, data = train\_SME)

# use step() to run forward selection

SME\_data.lm.step <- step(SME\_data.lm.null,

scope=list(lower=SME\_data.lm.null, upper=SME\_data.lm), direction =

"forward")

summary(SME\_data.lm.step)

Output:

A screenshot of a computer screen

Description automatically generated

1. **Backward R**

Code:

SME\_data.lm.step1 <- step(SME\_data.lm, direction = "backward")

summary(SME\_data.lm.step1)

Output:

A screenshot of a computer

Description automatically generated

1. **Stepwise R**

Code:

SME\_data.lm.step2 <- step(SME\_data.lm, direction = "both")

summary(SME\_data.lm.step2)

Output:

**A screenshot of a computer

Description automatically generated**

**List of Predictors:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Forward** | **Backward** | **Both** | **Exhaustive** |
| R.D.Spend | ✓ | ✓ | ✓ | ✓ |
| Marketing.Spend | ✓ | ✓ | ✓ | ✓ |
| Administration |  |  |  |  |
| State\_Newyork |  |  |  | ✓ |
| State\_California |  |  |  |  |

**MLR Comments:**

1. **R&D Spend**:
   * **Significance**: The p-value for this variable is extremely small (Pr(>|t|) ≈ 0), indicating that it is highly statistically significant. This means that the investment in R&D has a strong and positive association with profit.
   * **T-value**: The t-value for R&D Spend is 17.281, indicating strong statistical evidence against the null hypothesis (which would claim no relationship between R&D spend and profit).
2. **Marketing Spend**:
   * **Significance**: The p-value for Marketing Spend is 0.0974, which is marginally significant at the 10% level (indicated by the "." significance code). Though weaker than R&D Spend, this suggests some evidence that marketing spend has a positive impact on profit.
   * **T-value**: The t-value for Marketing Spend is 1.717, which is smaller than R&D Spend’s, but still shows some evidence of impact.

**Why these variables are chosen:**

The **exhaustive search** selection method, which evaluates all possible combinations of variables, identified **R&D Spend** and **Marketing Spend** as the most important predictors. This is reinforced by the results from the stepwise selection models, where only these two variables are consistently retained in all steps.

**Model Quality:**

* **Multiple R-squared (0.9628)**: This indicates that 96.28% of the variability in profit can be explained by the two predictors, R&D Spend and Marketing Spend.
* **Adjusted R-squared (0.9601)**: This is close to the multiple R-squared, confirming that the model doesn't overfit even after adjusting for the number of predictors.
* **F-statistic (349.7)**: The high F-statistic with a very low p-value suggests that the overall model is statistically significant.

In summary, **R&D Spend** is a very strong predictor of profit, while **Marketing Spend** provides additional (though weaker) explanatory power, which is why these two variables were selected as the best predictors in your model.

**PCA:**

Code:

head(SME\_data)

pca\_SME<-SME\_data[,1:3] ##removing categorical Variable

pca\_SME

pca<-prcomp(pca\_SME,scale. = TRUE)

summary(pca)

Output:

A screen shot of a computer

Description automatically generated

**Comments:**

PC1 and PC2 are the most important components, explaining 92.46% of the variance combined. These two components likely represent the key dimensions of variability in the dataset.

PC3 contributes minimally (7.54%) and may represent noise or less significant features. Therefore, it may be less valuable for further analysis.