Machine Learning Model

- Linear Regression
  - Understanding relationship between input and output variables
  - Input and output are numeric values

$$y= a_0+a_1x+ \varepsilon$$

Here,

Y= Dependent Variable (Target Variable)

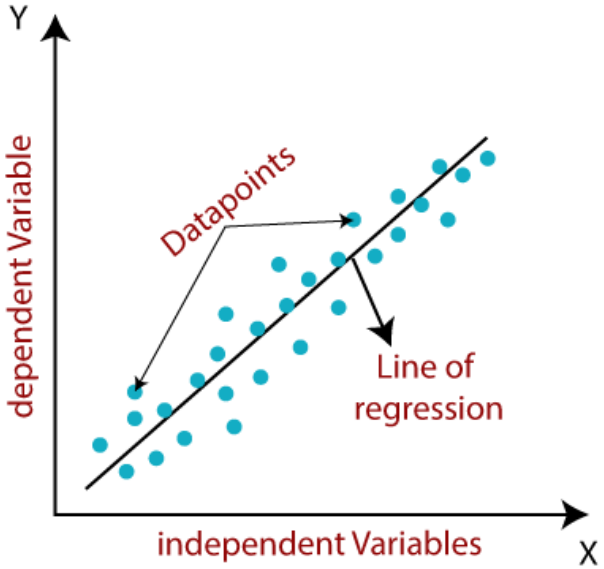
X= Independent Variable (predictor Variable)

a0= intercept of the line (Gives an additional degree of freedom)

a1 = Linear regression coefficient (scale factor to each input value).

ε = random error

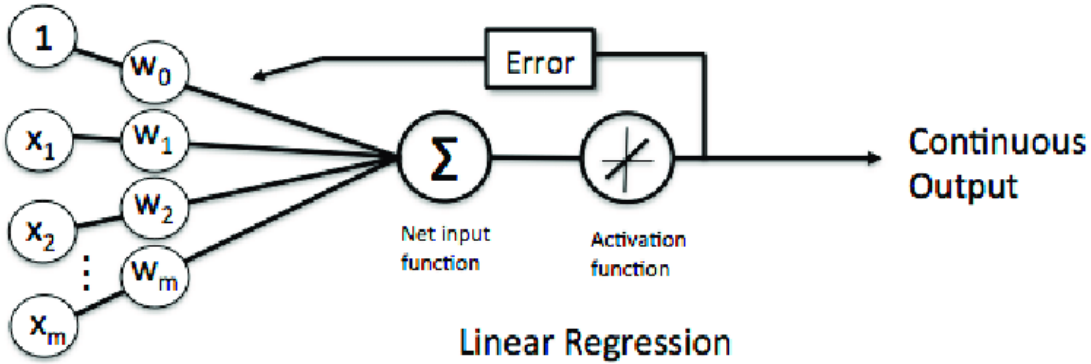
Source: https://www.javatpoint.com/linear-regression-in-machine-learning



Source: https://www.javatpoint.com/linear-regression-in-machine-learning

Machine Learning Model (cont'd)

- **Limitations**
  - Cannot predict discrete values
  - Sensitive to outliers
  - Prone to underfitting
- **Benefits**
  - Can predict continuous values
  - Linearity & simple implementation
  - Reduces overfitting by regularization



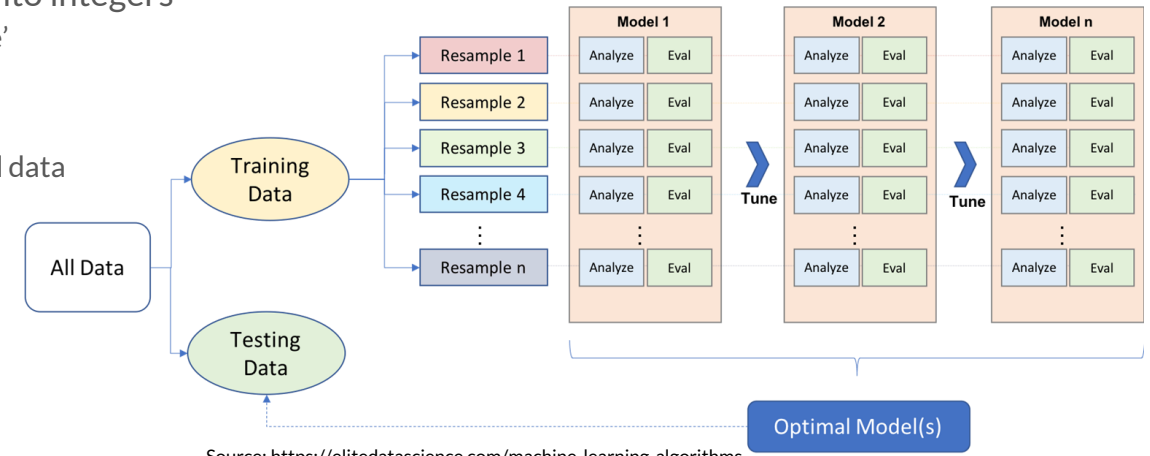
Source: https://www.researchgate.net/figure/Machine-learning-linear-regression_fig1_328948591

Initial Machine Learning Model

- Make dataset more compatible with model
  - Dropped ‘Date’ column
  - Dropped extreme outlier
- Convert ‘Holiday’ string values into integers
  - Grouped by ‘holiday’ and ‘none’
  - Converted labels into integers
- Scale the Data
  - Compared scaled and unscaled data

Initial Score

```
Training Score: 0.6994128803300881
Testing Score: 0.6565277870104217
```



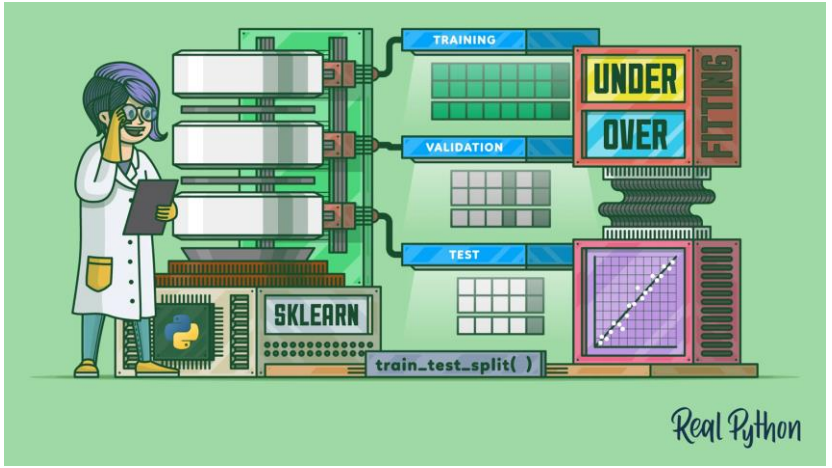
Source: https://elitedatascience.com/machine-learning-algorithms

Optimized Machine Learning Model

- Further Feature Engineering and Selection Experimenting
  - Can model's accuracy be improved?
  - Train model on new features.
  - Keep feature change if accuracy improved.
- Optimized R-squared Score: 65.85%
  - Included all initial features
  - Left each holiday value in place
  - Grouped days into weekday or weekend

Optimized Score

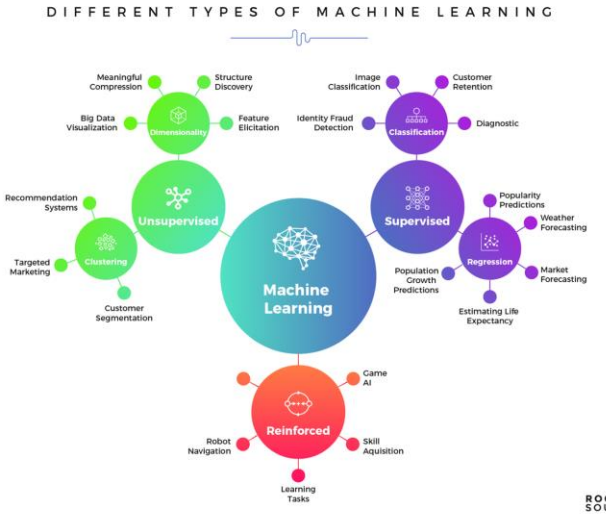
```
Training Score: 0.7007461735465028
Testing Score: 0.6585083431725288
```



Source: https://realpython.com/train-test-split-python-data/

What we would change

- Utilize a different Machine Learning model
- Minimizing the reduction of data rows
    - Condensing 40,000+ rows of data to ~1,000
- Collect more data



ROCKET SOURCE

Future Analysis

- Obtain additional records
  - From same dataset source
- Add additional combination features
  - New datasets
  - Comparing accessibility to public transportation
- Asking new questions
  - How does gas prices affect commuter habits?

Source: https://realpython.com/train-test-split-python-data/