Supervised learning- Regression

Parcours Progis

Etudes, Medias, communication, Marketing

Bahareh Afshinpour.

07.11.2024

Université Grenoble Alpes

Sciences Po Grenoble UGA

References

- **Book: AI for Marketing and Product Innovation,** Dr. A. K., Pradeep, Andrew Appel, and Stan Sthanunathan.

- **Book**: Machine Learning Implémentation en Python avec Scikit-learn (2e édition)- Virginie MATHIVET

- **Book:** Hwang, Yoon. 2019. *Hands-On Data Science for Marketing*. Packt Publishing, Ltd. Chapter 3, 128–156.

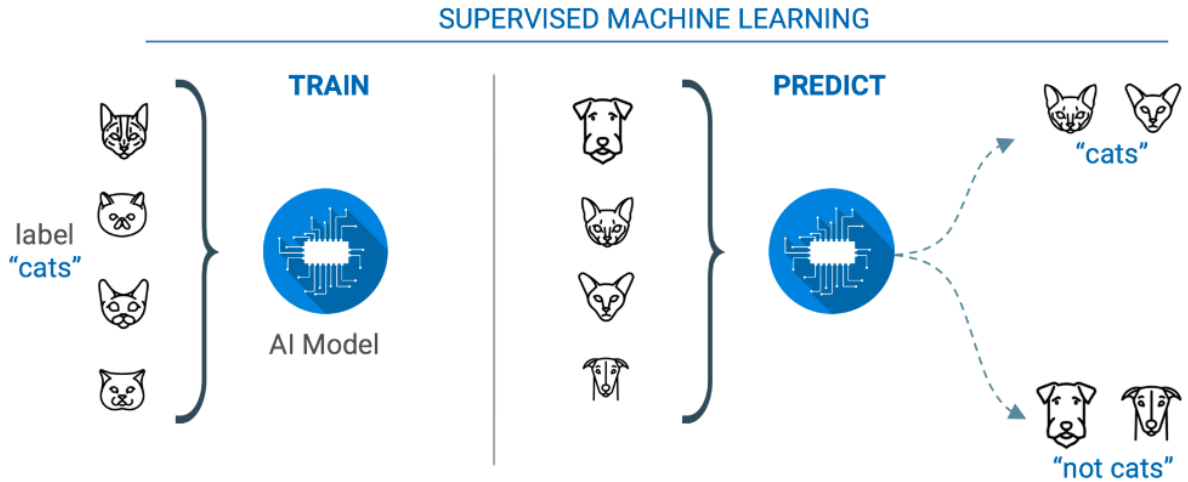
- https://www.youtube.com/watch?v=XA3OaoW86R8

Supervised learning

Supervised learning is a fundamental branch of machine learning where algorithms, learned patterns, and relationships in

labeled training data

to make **predictions** or **classify** new **unseen** data points.



https://abeyon.com/how-do-machines-learn/

- Marketing is still about reaching customers effectively, informing them, persuading them, motivating them, and ideally, bringing them back for more.
- Here, AI and ML algorithms are designed to execute some major tasks:

✓predict outcomes, for instance, by designing the algorithms to provide answers to questions such as

  ✓if someone discovers your product or service today, how likely is that person to sign up or make a purchase?

✓Which visitors are most likely to buy your product or service?

✓How much will a customer spend during his or her lifetime on your product or service?

Why do you need datasets?

- Machine learning algorithms learn from data. A machine learning algorithm identifies trends, and relationships, and makes predictions based on large volume of data given to training the models.

One of the major challenges faced by marketers today is the limited access **to customer data**. This is due to privacy regulations, increasing consumer concerns about data privacy, and the complexity of gathering and analyzing large volume of data.

What is a Supervised Learning Dataset?

- One of the fundamental steps in any machine learning project is importing the dataset.
- What is a DATASET?
  - A dataset in machine learning is a collection of instances (instance refers to a single row of data) that all share some common features and attributes.
  - Two kinds of datasets are required

    -**Training Dataset** — The data that is fed into the machine learning algorithm for training.

    -**Test Dataset** or Validation Dataset — The data that is used to evaluate and test that the machine learning model is interpreting accurately.

Dataset example (supervised)

buying	maint	doors	persons	lug_boot	safety	class
vhigh	med	2	more	med	high	acc
vhigh	med	2	more	big	low	unacc
vhigh	med	2	more	big	med	acc
vhigh	med	2	more	big	high	acc
vhigh	med	3	2	small	low	unacc
vhigh	med	3	2	small	med	unacc
vhigh	med	3	2	small	high	unacc
vhigh	med	3	2	med	low	unacc
vhigh	med	3	2	med	med	unacc
vhigh	med	3	2	med	high	unacc
vhigh	med	3	2	big	low	unacc
vhigh	med	3	2	big	med	unacc
vhigh	med	3	2	big	high	unacc
vhigh	med	3	4	small	low	unacc

buying	Feature	Categorical	buying price
maint	Feature	Categorical	price of the maintenance
doors	Feature	Categorical	number of doors
persons	Feature	Categorical	capacity in terms of persons to carry
lug_boot	Feature	Categorical	the size of luggage boot
safety	Feature	Categorical	estimated safety of the car
class	Target	Categorical	evaluation level (unacceptable, acceptable, good, very good)

https://www.kaggle.com/datasets/stealthtechnologies/car-evaluation-classification

How can we access the datasets?

https://blog.learnamp.com/why-asking-questions-is-the-key-to-business-success

Importing the Data Set in Machine Learning Using Sklearn datasets

- The scikit-learn library offers several datasets for machine learning. These datasets are available through sklearn.datasets and fall into different categories

  - `load_iris` : Iris flower dataset (classification)
  - `load_digits` : Handwritten digits dataset (classification)
  - `load_wine` : Wine dataset (classification)
  - `load_breast_cancer` : Breast cancer dataset (classification)
  - `load_diabetes` : Diabetes dataset (regression)
  - `load_linnerud` : Linnerud dataset (multivariate regression)

-Importing Scikit-Learn

make sure you have Scikit-Learn installed. If not, you can install it using:

```
pip install scikit-learn
```

-Loading Sample Datasets

To load a sample dataset, you can use the following code:

```
from sklearn import datasets
# Load the iris dataset
iris = datasets.load_iris()
 # Access the features and target variable
X = iris.data # Features
y = iris.target # Target variable
```

Importing the Data Set in Machine Learning Using Sklearn datasets

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
plt.rcParams['figure.figsize'] = [12, 4]
from sklearn import datasets

# Load the diabetes dataset
diabetes = datasets.load_diabetes(as_frame=True)
print(diabetes.DESCR)
```

```
**Data Set Characteristics:**

  :Number of Instances: 442

  :Number of Attributes: First 10 columns are numeric predictive values

  :Target: Column 11 is a quantitative measure of disease progression one ye

  :Attribute Information:
      - age     age in years
      - sex
      - bmi     body mass index
      - bp      average blood pressure
      - s1      tc, T-Cells (a type of white blood cells)
      - s2      ldl, low-density lipoproteins
      - s3      hdl, high-density lipoproteins
      - s4      tch, thyroid stimulating hormone
      - s5      ltg, lamotrigine
      - s6      glu, blood sugar level
```

Different forms of data

Grouping by color is one way to organize categorical data.

- Categorical data:
  - A categorical characteristic is an attribute that can take one of a limited and usually fixed number of possible values based on a qualitative property. For example, hair color (black, orange, blond)
  - Example in marketing: understanding customer segments through categorical data allows for targeted advertising and personalized marketing campaigns.
- Numerical data:
  - If an entity represents a characteristic measured in numbers, it is called a **numerical entity**.
  - **Discrete data**: countable items. For example Shoe sizes.
  - **Continuous data:** can take any value within a range, such as temperatures or distances.

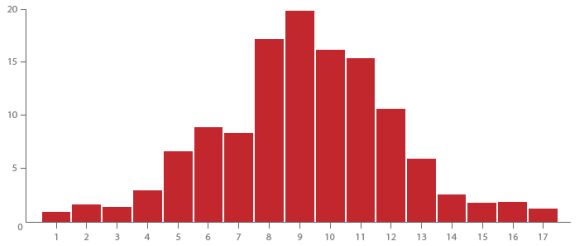
Types de données

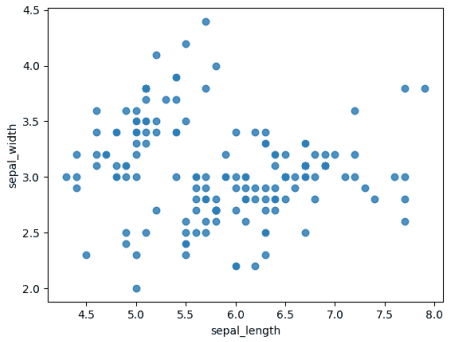
- Pour chaque variable du jeu de données, deux informations devront être produites :
  - une analyse statistique, sous forme textuelle ou tabulaire.
  - une représentation graphique des données (visualisation des données)

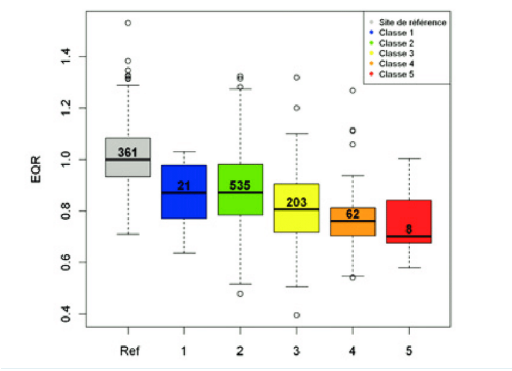
- Les outils à utiliser sont cependant différents selon le type de la variable (catégorielle ou numérique)

Une représentation graphique des données

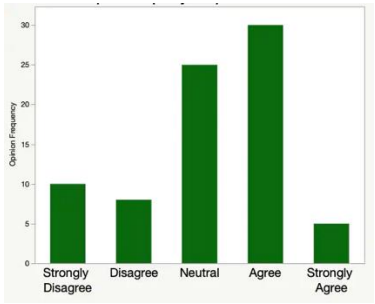
- Graphiques sur données numériques
  - Histogramme, Nuage de points(Scatter plot), Boites a moustaches (box plot)

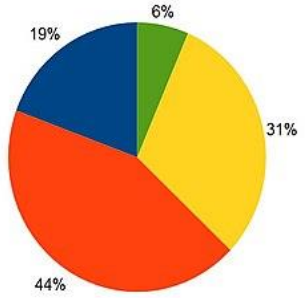






- Graphiques sur données catégorielles
  - Les diagrammes à barres, les diagrammes circulaires (aussi appelés « en camembert »)





Importing the Data Set in Machine Learning Using Scikit-Learn from CSV file

- **Loading CSV Datasets**

For datasets in CSV format, you can use libraries like Pandas to load the data and then convert it to NumPy arrays for further processing with Scikit-Learn:

```
data = pd.read_csv('./WA_Fn-UseC_-Marketing-Customer-Value-Analysis.csv', encoding='latin1')
data.columns
```

```
Index(['Customer', 'State', 'Customer Lifetime Value', 'Response', 'Coverage',
       'Education', 'Effective To Date', 'EmploymentStatus', 'Gender',
       'Income', 'Location Code', 'Marital Status', 'Monthly Premium Auto',
       'Months Since Last Claim', 'Months Since Policy Inception',
       'Number of Open Complaints', 'Number of Policies', 'Policy Type',
       'Policy', 'Renew Offer Type', 'Sales Channel', 'Total Claim Amount',
       'Vehicle Class', 'Vehicle Size'],
      dtype='object')
```

Data cleansing (data scrubbing)

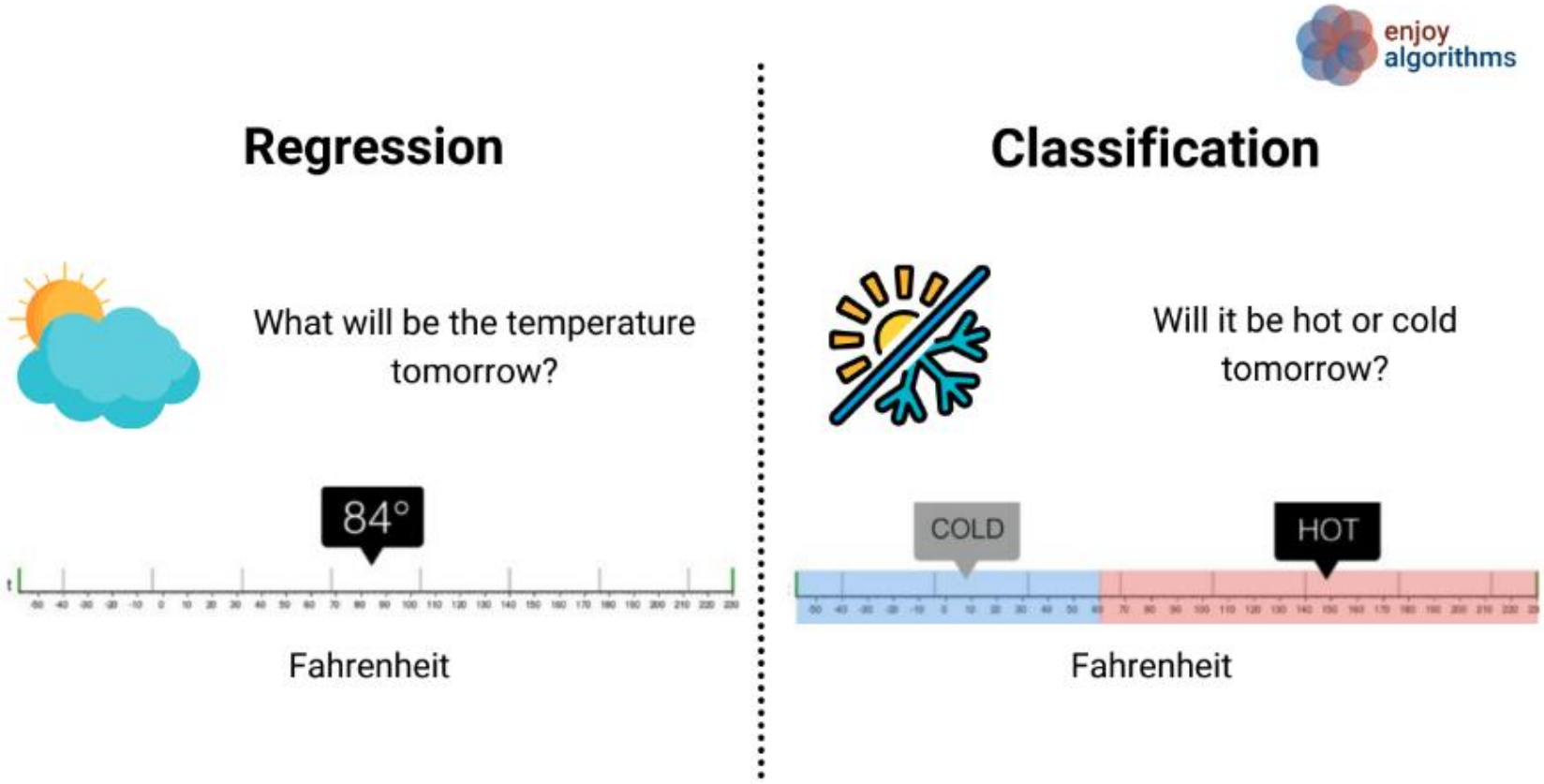
- data cleansing is the process of finding and correcting or deleting irrelevant, missing, duplicate, or otherwise useless data from a datset.
- This is a necessary step designed to purify data so that algorithms can work faster and make more accurate predictions.

-Reasons for the corruption of data are :
user error, dummy data, and workarounds.
-Data cleansing functions may include the enhancement, harmonization, and standardization of data.

*To perform data cleansing, all incorrect, incomplete, and irrelevant data must be found, then either **replaced**, **removed**, or **modified**.*

https://www.dataversity.net/what-is-data-cleansing

Supervised Learning



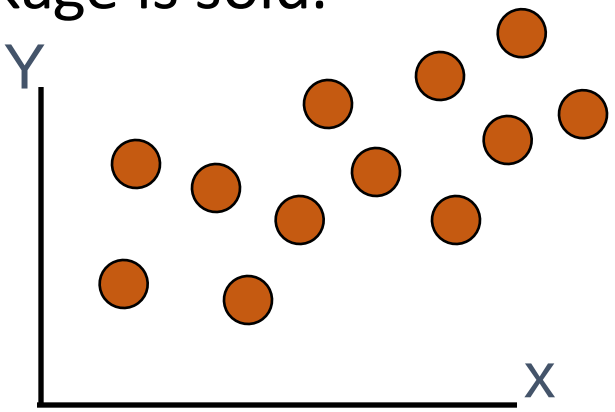
https://www.enjoyalgorithms.com/blogs/classification-and-regression-in-machine-learning

Regression

- It can help to invistigate the relation between variables.
- We commonly use three major types of regression:
  - Linear regression
  - Polynomial regression
  - And Logistic regression

Regression

- Imagine we collect the price of a package of pumpkins on different days of the year.
- the x value of each point is **the day of the year** the package is sold.
- And the Y value is **the price** of the package.



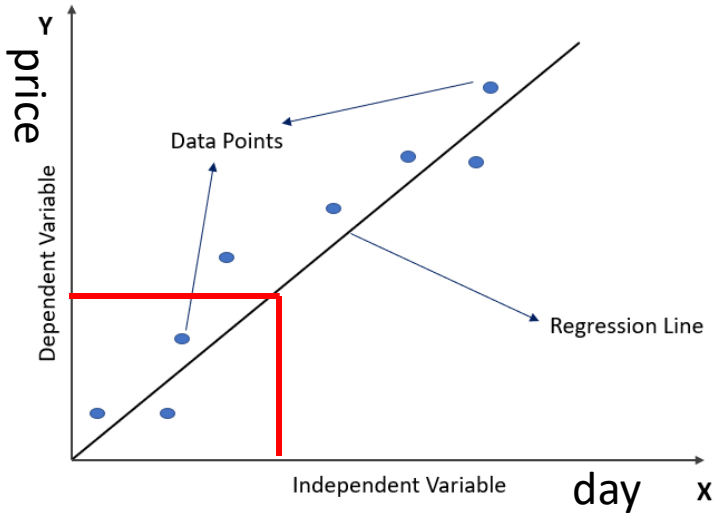
- You can use regression to find a **mathematical formula** that represents the general trend of the data.
- This formula **encodes the relationship** between our x and y variables
- And enables us to **predict y for any given value of x**.

Linear Regression

- It uses a straight line to approximate the trend of our data points.
- **y = mx + b**, where m is the slope of the line, b is the y-intercept, and x is the independent variable.
- A slope of 2 means that every 1-unit change in X yields a 2-unit change in Y.
- In pumpkin dataset, we want to predict

the price of a package of pumpkins for the

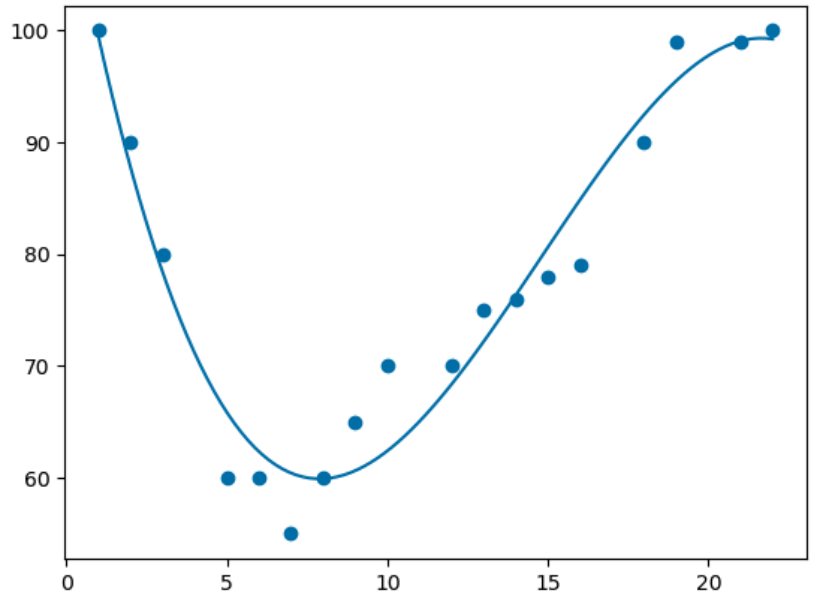
Particular day of the year.



https://wisdomml.in/linear-regression-in-machine-learning-a-comprehensive-guide/

Polynomial regression

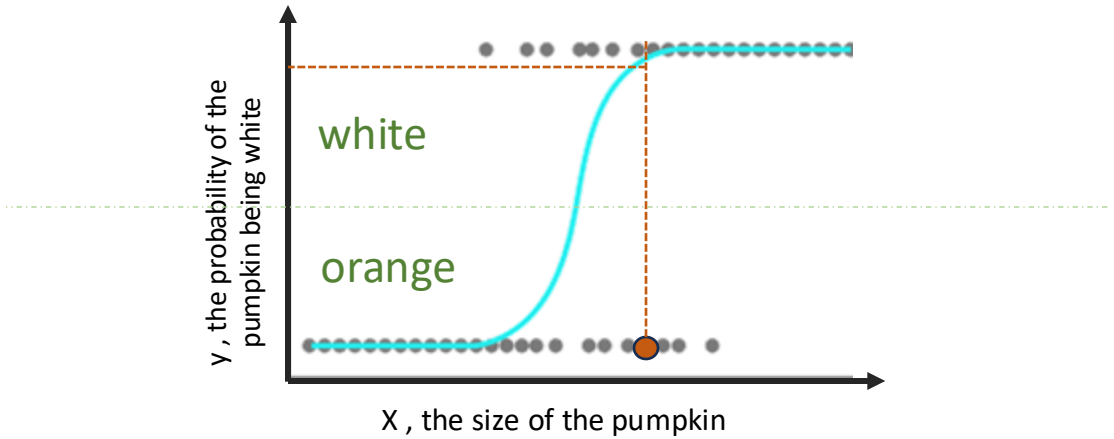
- In some situations, we can consider polynomial regression which is really just an extension of linear regression that uses a curve to represent a relationship between variables.
- We introduce a new term in our formula that include the square of x variable.



Logistic regression

- We want to predict whether the color of a pumpkin is orange or white based on its size.
- The value we want to predict is a category either orange or white.

When we want to predict a category logistic regression is a great method



Exemple de cas pratique

- La regression est utilisée dans de nombreux cas potentiels :
  - Prédire le prix d'un produit ou service en fonction de ses caractéristiques, comme le prix d'un appartement en fonction de sa taille, le nombre de chambres , etc.
  - Évaluer les risques d'un évènement en fonction d'autres évènements ou informations.
  - Estimez la qualité d'une production en fonction des données physiques du processus (température, pression, humidité…).
  - ….

Drivers Behind Marketing Engagement

- Regression analysis allows marketers to explore the relationships between independent variables, such as advertising spend, social media presence, customer demographics, and the dependent variable of marketing engagement.

- By analyzing the statistical significance and coefficients of these variables, marketers can identify the key factors that influence engagement and gain insights into their relative impact.

Book: Hwang, Yoon. 2019. *Hands-On Data Science for Marketing*. Packt Publishing, Ltd. Chapter 3, 128–156.

Entrainement et Évaluation des modèles

- Avec Scikit-Learn, le processus sera toujours le même :
  1. Créer un modèle.
  2. Faire l'apprentissage grâce à **fit**, en lui fournissant les données X et y d'apprentissage.
  3. Prédire des résultats sur le dataset souhaité grâce à **predict** .
  4. Appler les différentes métriques souhaitées (presentes dans **sklearn.metrics**) avec en paramètres les résultats attendus et les données prédites.

En termes de code …

In [11]:
```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
```

In [12]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, t
est_size=0.4)
```

In [13]:
```
lm = LinearRegression()
lm.fit(X_train,y_train)
```

In [14]:
```
predictions = lm.predict(X_test)
```

In [16]:
```
from sklearn import metrics

print('MAE:', metrics.mean_absolute_error(y_test, predictio
ns))
print('MSE:', metrics.mean_squared_error(y_test, prediction
s))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, p
redictions)))
```

```
MAE: 3.53544941908
MSE: 20.8892997114
RMSE: 4.57048134351
```

Accuracy of linear regression model

- Numerous elements, including the quantity and quality of the data, the selection of independent variables, and the assumptions made regarding the relationships between the variables, affect how accurate the linear regression model is.

- It's crucial to choose the independent variables carefully and perform exploratory data analysis to make sure the assumptions are met in order to guarantee the accuracy of the model.

Accuracy of linear regression model

- After you check for the validity of using linear regression, you will now need to ask yourself how to train this model.
- So, first we have to define an error function or an evaluation metric to check the performance of our model, which can be one of the following:
  - **Mean Absolute Error (MAE):** we calculate the average absolute difference between the actual values and the predicted values.
  - **Root Mean Square Error (RMSE):** RMSE calculates the square root average of the sum of the squared difference between the actual and the predicted values.
  - ……

$$\text{MAE} = \underbrace{\frac{1}{n}}_{\text{test set}} \sum_{i=1}^{n} \underbrace{|y_i}_{\text{predicted vaue}} - \underbrace{\hat{y}_i|}_{\text{actual value}} \qquad \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2}$$

Accuracy of linear regression model

- There are several metrics used to determine the accuracy of a linear regression model, including
  - mean squared error (MSE),
  - root mean squared error (RMSE), and
  - R-squared
- These metrics measure the difference between the predicted values and the actual values.

END