Big Data

YAKA 2023 Team

Copyright

Rules

- You must have downloaded your copy from the Assistants' Intranet <https://intra.assistants.epita.fr>.
- This document is strictly personal and must **not** be passed on to someone else.
- Non-compliance with these rules can lead to severe sanctions.

Introduction

- The Data: stock for apple company

Composition

- `Date`
- `High`: highest price in the day
- `Low`: Lowest price in the day
- `Open`: Price at oppening
- `Close`: Price at Closing
- `Colume`: the number of trades in the day
- `Adj Close`: the closing price after accountment for corporate actions
- `company_name`

Goals

Find the value of `Adj Close` for the next day

First step: Cleaning Data so removing useless column (company_name, High, Low, Date)

Second Step: Add Column `Adj Close of Tomorrow` wich contain the value of Adj Close for tomorrow

Third Step: Separate the Data Set in two part one for training the other for testing

Our Model: Linear Regression

We choose the column to put in input:

- Open
- Close
- Volume
- Adj Close

Then we train our model with the vector assembler. And we test it with our test data to check if the prediction are correct.

The average precision that we got is: 0,022.

Evalution of our Model: Regression evaluator

To evaluate the result of our model we want to use the Regression evaluator, so calculate the Root Mean Square Error.

To be able to know how good it is, we want first to have a baseline model, to know if we do better than him. The baseline that we choose is the average of `adjusted close of Tomorrow` from our test database.

With this we got 23,58.

Then we create an other Regression evaluator based on our prediction.

With this we got 1,59.

We conclude that our model is doing better than the average, so our model is working and validated.

Questions

Any questions?