## Assignment 8

## Tips

* Question 1:

average tip:

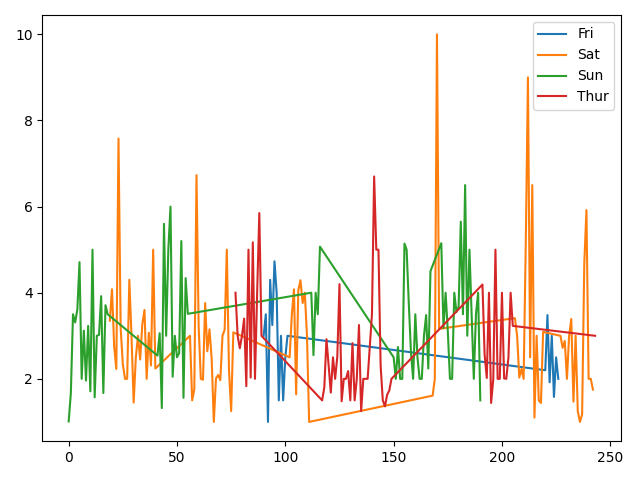
1. lunch 16.41%
2. dinner 15.95%

* Question 2:

average tip

1. Friday 16.99%
2. Saturday 15.32%
3. Sunday 16.69%
4. Thursday 16.13%

* Question 3: the highest tips is Friday lunch
* Question 4: correlation between is 0.68
* Question 5: correlation between is 0.49, so there is a positive relationship between tips and group size.
* Question 6: percentage of people are smoking is 38.11%
* Question 7: plot



The tips are not increasing with time in each day.

* Question 8: the correlation between tip and smoker is 0.0059, so the tip amount has no relationship with smoker.

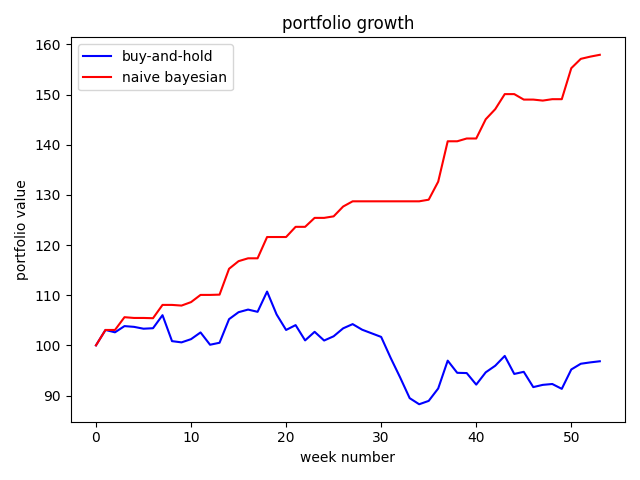
## Naïve Bayesian

* Question 1: accuracy for year 2 is 94.34%
* Question 2: confusion matrix

[[29 0]

[3 21]]

* Question 3: true positive rate: 100% true negative rate 87.5%
* Question 4: plot, we can find in this chart that Gaussian naïve Bayesian give us a large amount.



## Shapley feature explanations

* Question 1: table

|  |  |  |  |
| --- | --- | --- | --- |
|  | Logistic regression | kNN (k=7) | Linear model |
|  | 32.08% | 28.3% | 3.77% |
|  | -1.89% | -13.21% | 1.89% |

From the table above, we can find that if we remove the standard deviation column the accuracy is much higher. And if we remove the mean return column the accuracy is lower than use two columns to predict. Also, we can find the mean return column is more important than the volatility. And, the linear model may not cause a big change of accuracy.

* Question 2: table

|  |  |  |  |
| --- | --- | --- | --- |
| Flower | Versicolor | Setosa | Virginica |
| Sepal length | -1.33% | 0.0% | 0.0% |
| Sepal width | 8.0% | 0.0% | -1.33% |
| Petal length | 0.0% | 0.0% | -1.33% |
| Petal width | 1.33% | 0.0% | -1.33% |

From the table above, we can find that flower setosa, may not affect by those four features. And, three features cause the same result of virginica flower. Sepal length and petal length are the most irrelevant features.