Holden Herrell
14 March 2020

IST 718: Lab 3

Introduction

Machine learning has become much more prevalent in today's research efforts. Machine learning can be powerful; however, it can also require substantial computing resources if models are not tuned and parameterized properly. Zalando Research is seeking to identify the most accurate and efficient method to identify articles of clothing from images using machine learning. They have narrowed their desired models down to either a Neural Network Model, Naïve Bayes Model, or Random Forest Decision Tree Model. To help them make their final decision, they have sponsored a study of these three models on clothing image data.

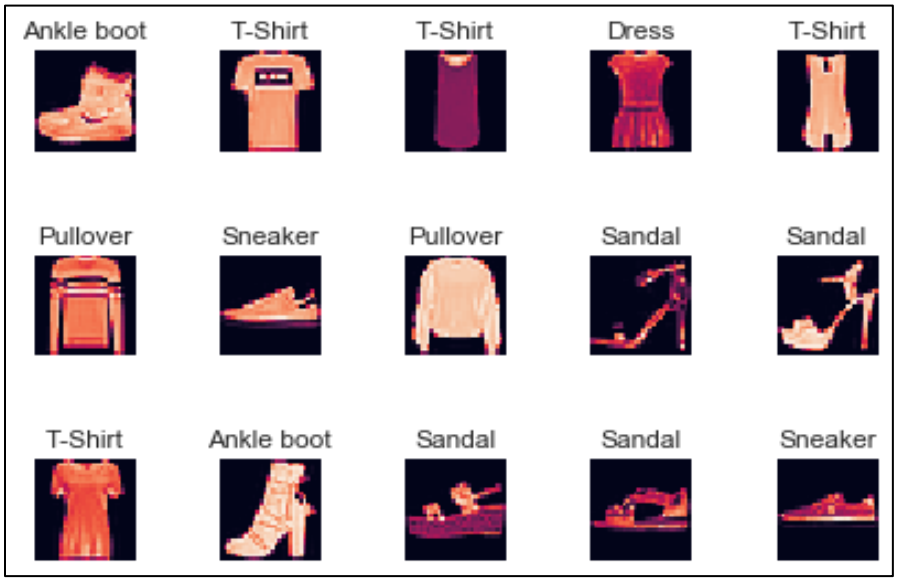
About the Data

The data for this study is sourced from Kaggle's Fashion MNIST dataset (https://www.kaggle.com/zalando-research/fashionmnist). The data consists of a clothing item label and pixel values of the images where each row represents an image of an item of clothing. This clothing types labeled within the data are as follows:

Label	Description	Label	Description
0	T-shirt	5	Sandal
1	Trouser	6	Shirt
2	Pullover	7	Sneaker
3	Dress	8	Bag
4	Coat	9	Ankle boot

A sample of the data rendered as images can be seen in Figure 1 (below).

Figure 1



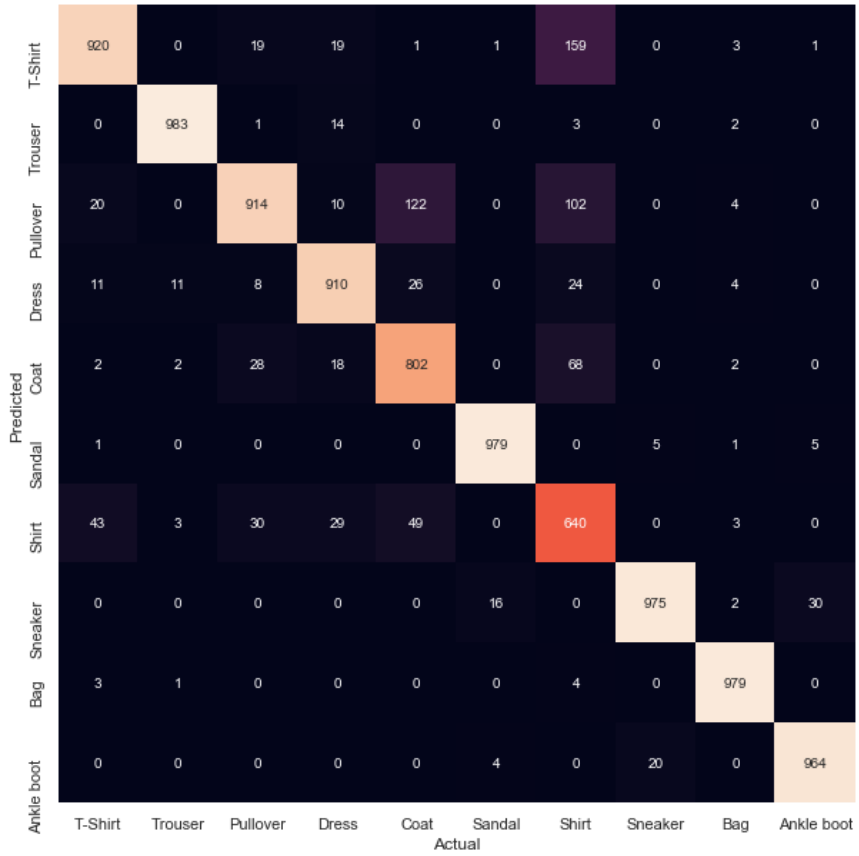
Neural Network

The first model tested for prediction accuracy and efficiency was the Neural Network Model. The Neural Network took 13min 9sec to train and was around 91% accurate at predicting the clothing images. The output statistics can be seen in Figure 2 and the confusion matrix in Figure 3.

Figure 2

	precision	recall	f1-score	support
T-Shirt	0.82	0.92	0.87	1000
Trouser	0.98	0.98	0.98	1000
Pullover	0.78	0.91	0.84	1000
Dress	0.92	0.91	0.91	1000
Coat	0.87	0.80	0.83	1000
Sandal	0.99	0.98	0.98	1000
Shirt	0.80	0.64	0.71	1000
Sneaker	0.95	0.97	0.96	1000
Bag	0.99	0.98	0.99	1000
Ankle boot	0.98	0.96	0.97	1000
accuracy			0.91	10000
macro avg	0.91	0.91	0.91	10000
weighted avg	0.91	0.91	0.91	10000

Figure 3



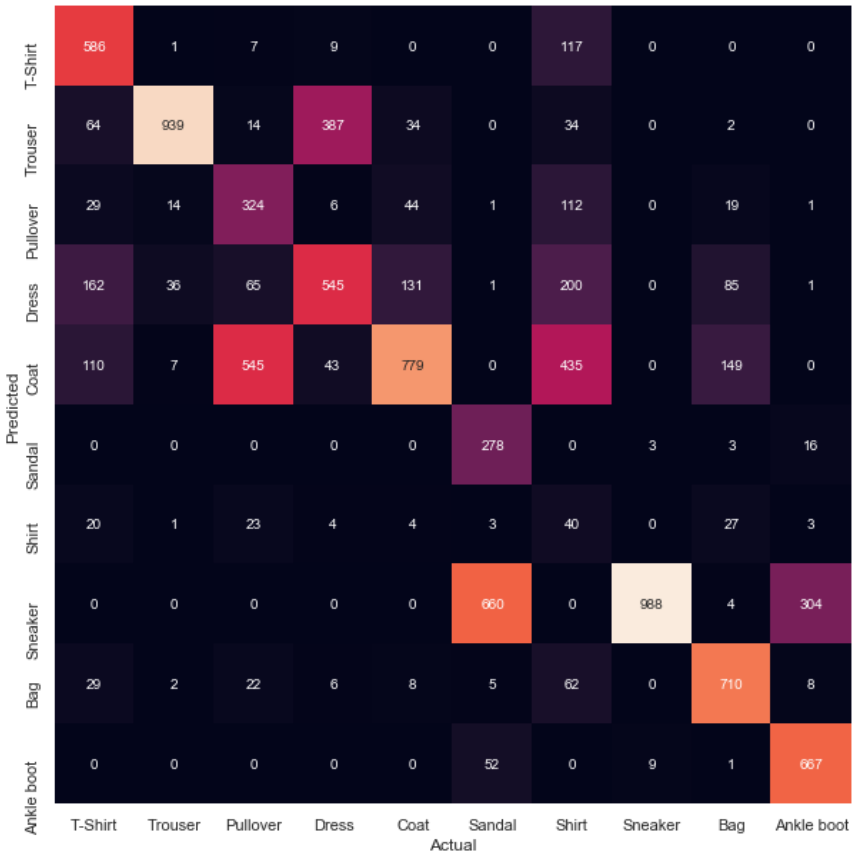
Naïve Bayes

The second model tested for tested was the Naïve Bayes Model. The Gaussian version of the Naïve Bayes Model was used for the purpose of this study. The model took 675ms to train and was around 60% accurate. The output statistics can be seen in Figure 4 and the confusion matrix in Figure 5.

Figure 4

	precision	recall	f1-score	support
T-Shirt	0.81	0.59	0.68	1000
Trouser	0.64	0.94	0.76	1000
Pullover	0.59	0.32	0.42	1000
Dress	0.44	0.55	0.49	1000
Coat	0.38	0.78	0.51	1000
Sandal	0.93	0.28	0.43	1000
Shirt	0.32	0.04	0.07	1000
Sneaker	0.51	0.99	0.67	1000
Bag	0.83	0.71	0.77	1000
Ankle boot	0.91	0.67	0.77	1000
accuracy			0.59	10000
macro avg	0.64	0.59	0.56	10000
weighted avg	0.64	0.59	0.56	10000

Figure 5



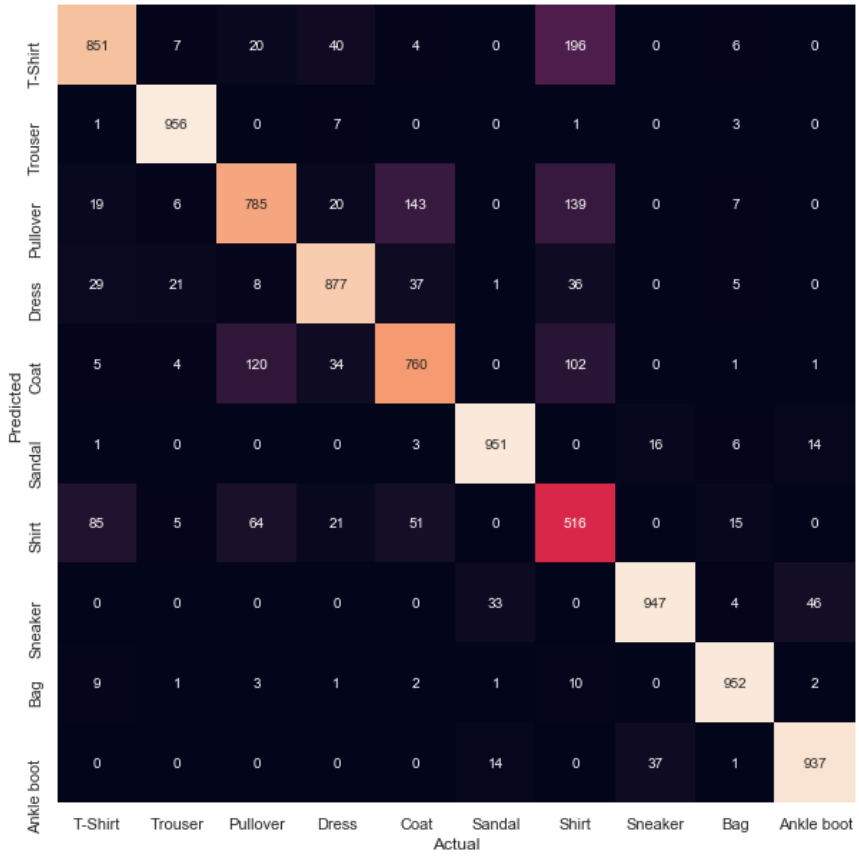
Random Forest

The final model tested for this study was the Random Forest Decision Tree Model. This model took 8.12sec to train and was around 85% accurate. The output statistics and confusion matrix can be seen in Figures 6 and 7 respectively.

Figure 6

	precision	recall	f1-score	support
T-Shirt	0.76	0.85	0.80	1000
Trouser	0.99	0.96	0.97	1000
Pullover	0.70	0.79	0.74	1000
Dress	0.86	0.88	0.87	1000
Coat	0.74	0.76	0.75	1000
Sandal	0.96	0.95	0.96	1000
Shirt	0.68	0.52	0.59	1000
Sneaker	0.92	0.95	0.93	1000
Bag	0.97	0.95	0.96	1000
Ankle boot	0.95	0.94	0.94	1000
accuracy			0.85	10000
macro avg	0.85	0.85	0.85	10000
weighted avg	0.85	0.85	0.85	10000

Figure 7



Comparing the Models

When comparing the three models, the Neural Network is the most accurate, followed rather closely by the Random Forest Model. However, both models seem substantially more accurate than the (Gaussian) Naïve Bayes model. In terms of efficiency, the Naïve Bayes model took the least amount of time to train, but reduced training time seems to lead to reduced accuracy. The Random Forest model was ~97 times faster to train than the Neural Network (8.12sec vs 13min 9sec), and only sacrificed 6% accuracy. It would depend primarily on Zalando's preference between precision and performance to say which model best suits their needs. The Neural Network is the most accurate, however the Random Forest Model was the most efficient (see below).

Model	Precision	Recall	F1-score	Accuracy	Support	Train Time (sec)
Neural Network	0.91	0.91	0.91	0.91	10000	789.000
Naïve Bayes	0.64	0.59	0.56	0.59	10000	000.675
Random Forest	0.85	0.85	0.85	0.85	10000	008.120

Conclusions

Zalando Research is seeking to identify the most accurate and efficient method to identify articles of clothing from images using machine learning. They are trying to decide between a Neural Network Model, Naïve Bayes Model, or Random Forest Decision Tree Model. This study found that if they value solely the highest accuracy, the Neural Network Model best suits their needs. However, this model is the least performant in terms of computing time required to train the model. The Random Forest Model (although slightly less accurate) offered a much more efficient option with a fraction of the compute time. If Zalando is comfortable with sacrificing a small percentage of accuracy for an almost 100x improvement in train time (compared to the Neural Network), then the Random Forest Model would be best.