

Classy NOT Trashy

Garbage Classification using Apache Spark



05.10.2021

Authors: Akshata Moharir, Aditya Deshpande, Camilla Bendetti

Problem Statement

- 75% of garbage can be recycled but instead ends up in landfills.
- Sorting garbage is costly and time consuming.

However,

Sorting garbage increases recycling therefore reducing the amount of waste produced and in turn the amount of waste in landfills.

Our Solution

- Find a way to sort garbage that will be more cost and time effective than a manual sorting system








- We have developed an image classifier using the transfer learning algorithm to predict the type of garbage with **77%** accuracy.

Data & Technology Stack

- The image data was collected from [Kaggle.com](https://www.kaggle.com) and is comprised of six different classes of garbage all captured in jpg format
 - Cardboard, Plastic, Metal, Glass, Paper, and Trash
 - Between 100 and 600 images for each class
- We leveraged Apache Spark and its deep learning library, sparkdl, to perform the object detection with python as the programming language

```
display(image_df)
```

	image
1	
2	
3	
4	
5	

	
43	
44	
45	

Feature Engineering and Model Development : Using Inception V3 transfer learning model

FEATURE ENGINEERING

Our current framework uses the transfer learning concept to shortlist the relevant features required for the object classification model, we have used the Inception V3 existing model to drive the relevant set of features.

MODEL DEVELOPMENT

Logistic Regression Model

Accuracy 0.7698



Random Forest Model

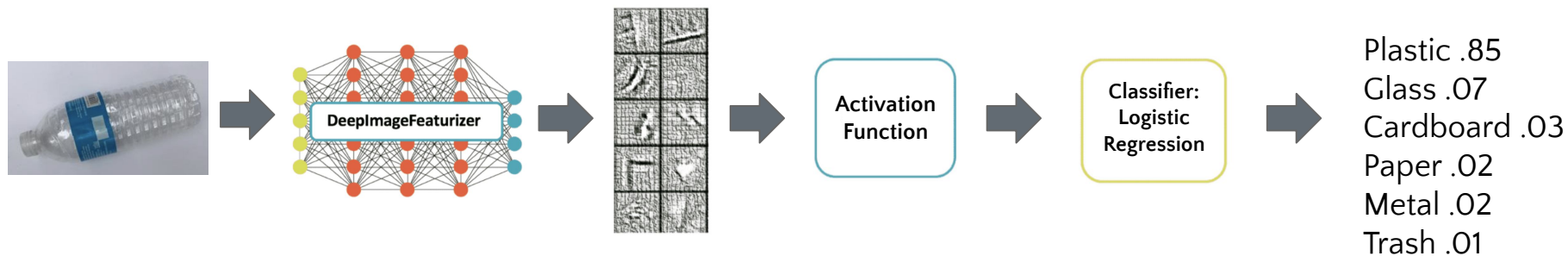
Accuracy 0.6128

Decision Tree Model

Accuracy 0.4679

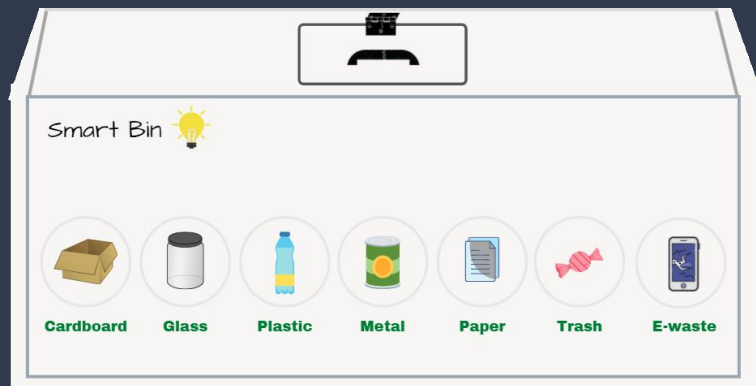
image label	features	rawPrediction	probability prediction
[dbfs:/tmp/garbag... 4	[0.75031846761703...	[-0.2359312291565...	[0.09194024500730... 4.0
[dbfs:/tmp/garbag... 4	[0.0,0.0310294590...	[-1.1270310344171...	[0.02146830652470... 4.0
[dbfs:/tmp/garbag... 4	[0.46827483177185...	[-1.3118962175287...	[0.01890508300402... 4.0
[dbfs:/tmp/garbag... 4	[0.0,0.0,0.0,0.04...	[-0.5677108304337...	[0.04509343539443... 4.0
[dbfs:/tmp/garbag... 4	[0.0,0.0867210030...	[-0.5421447372914...	[0.04051402396789... 4.0
[dbfs:/tmp/garbag... 1	[0.0,0.0,0.0,1.20...	[-0.9848781882571...	[0.03297343822379... 1.0
[dbfs:/tmp/garbag... 4	[0.0,0.3163996636...	[-1.9670428445055...	[0.01482026514289... 4.0
[dbfs:/tmp/garbag... 5	[0.0,0.9734503030...	[-0.6101738954507...	[0.06025791135065... 4.0
[dbfs:/tmp/garbag... 4	[0.76288890838623...	[-1.6137097940157...	[0.02137695177214... 4.0
[dbfs:/tmp/garbag... 4	[1.63313889503479...	[-1.4626518224919...	[0.02232426374705... 4.0
[dbfs:/tmp/garbag... 4	[0.30167287588119...	[-1.3782271559644...	[0.00804064015650... 4.0
[dbfs:/tmp/garbag... 0	[0.0,0.0,0.0,0.52...	[0.56483293626922...	[0.20705483667768... 2.0
[dbfs:/tmp/garbag... 4	[1.32918953895568...	[-0.8867917103480...	[0.03644450370430... 4.0
[dbfs:/tmp/garbag... 4	[0.0,0.0,0.0,0.0,...	[-0.4885254631122...	[0.04246145380344... 4.0
[dbfs:/tmp/garbag... 4	[0.0,0.0,0.036602...	[-1.4730457229166...	[0.00722668635879... 4.0
[dbfs:/tmp/garbag... 5	[0.27335953712463...	[-1.3924344882497...	[0.02392043707473... 5.0
[dbfs:/tmp/garbag... 5	[0.0,0.6644692420...	[-0.1894652830923...	[0.08618694118632... 5.0
[dbfs:/tmp/garbag... 1	[0.16303347051143...	[-0.0693338820250...	[0.08023097981569... 2.0

Deep Learning Framework for Garbage Classification



Results and Conclusion

Final Model: Logistic Regression



We compared the results from the three different models and found out that the Logistic Regression model outperformed in terms *accuracy*.

We believe that garbage classification at source is easiest and most cost effective. Therefore, separating the garbage at its source is much better than collecting all the garbage and then sorting at a central facility by hand.

In the future, we intend to develop the model in near - real time and deploy it by leveraging other big data libraries like BigDL and Apache Zoo for improving the accuracy of the model .

For more information please view our [Github](#).

Sources

Academic Sources:

N.-A-A.;Ahsan,M.;Based, M.A.; Haider, J.; Kowalski, M.
COVID-19 Detection from Chest X-ray Images Using
Feature Fusion and Deep Learning. *Sensors* **2021**, *21*, 1480.
<https://doi.org/10.3390/s21041480>

Image classification model based on spark and CNN
Jiangfeng Xu, Shenyue Ma
MATEC Web Conf. 189 03012 (2018)
DOI: 10.1051/matecconf/201818903012

Data and Code Sources:

<https://www.kaggle.com/asdasdasdas/garbage-classification>

<https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bfcf/5669198905533692/3647723071348946/3983381308530741/latest.html>

Images:

<https://duaneleffel.com/blog/2019/3/7/tips-for-recycling>

<https://www.kdnuggets.com/2017/09/databricks-vision-making-deep-learning-simple.html>

https://www.freepik.com/free-vector/garbage-sorting-set_13146308.htm

Other:

<https://www.holdenbags.com/blog/why-we-need-to-sort-garbage#:~:text=Sorting%20your%20garbage%20helps%20reduce,services%20and%20even%20litter%20collection>