Is the E-Mail a SPAM or a HAM?

A simple Decision Tree Based classifier to identify if a given mail is spam or ham

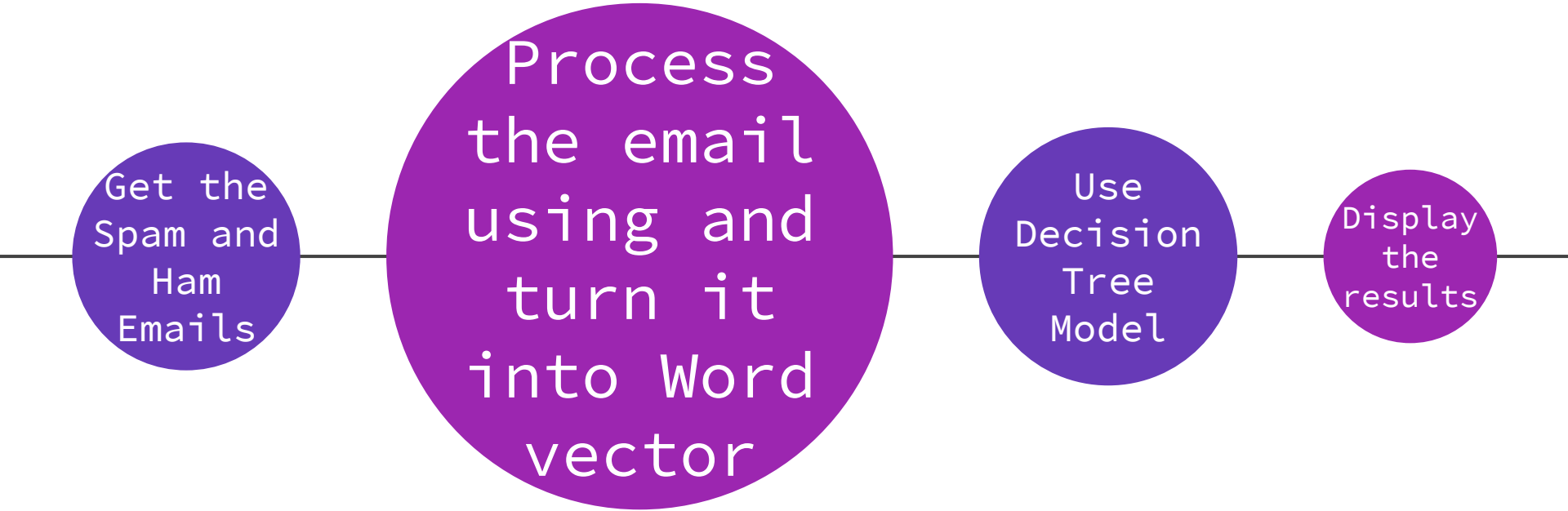
-BY Vikraant Pai
Roll No. D005
SAP ID: 70271019005

About The Problem

---

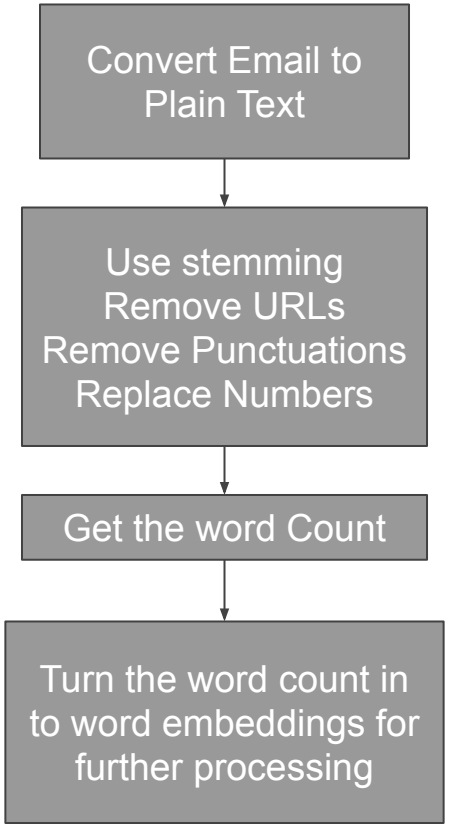
1. The problem that is being solved here is finding the right decision tree model that can exam whether the mail is spam or a ham
2. The method that is used to solve this problem is to first extract information from the mail to find out the words in the mail and then fit a decision tree model
3. Using this model it is tested on a test sample for precision and recall and compared with logistic regression

Steps Taken to build the model



Extracting the features from the mail

Email is converted to plain text and then word vector using Natural language Toolkit for stemming and based on the size of the vocabulary for the corpus, it is converted to word embeddings for further processing



Using the custom transformer to transform the words into Word embeddings

```
[27] from sklearn.pipeline import Pipeline

     preprocess_pipeline = Pipeline([
         ("email_to_wordcount", EmailToWordCounterTransformer()),
         ("wordcount_to_vector", WordCounterToVectorTransformer()),
     ])

     X_train_transformed = preprocess_pipeline.fit_transform(X_train)
```

Using the custom transformer to transform the words into Word embeddings

```
[112] from sklearn.tree import DecisionTreeClassifier
      dtc = DecisionTreeClassifier( max_depth=5,max_leaf_nodes=5, criterion="gini")
      dtc.fit(X_train_transformed, y_train)
      score = cross_val_score(dtc, X_train_transformed, y_train, cv=3, verbose=3)
```

```
[CV]  ................................................................
[CV] ................................. , score=0.931, total=   0.0s
[CV]  ................................................................
[CV] ................................. , score=0.940, total=   0.0s
[CV]  ................................................................
[CV] ................................. , score=0.929, total=   0.0s
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done   1 out of   1 | elapsed:    0.0s remaining:    0.0s
[Parallel(n_jobs=1)]: Done   2 out of   2 | elapsed:    0.0s remaining:    0.0s
[Parallel(n_jobs=1)]: Done   3 out of   3 | elapsed:    0.1s finished
```

```
print(f"Mean cross validation score {score.mean()}")
```

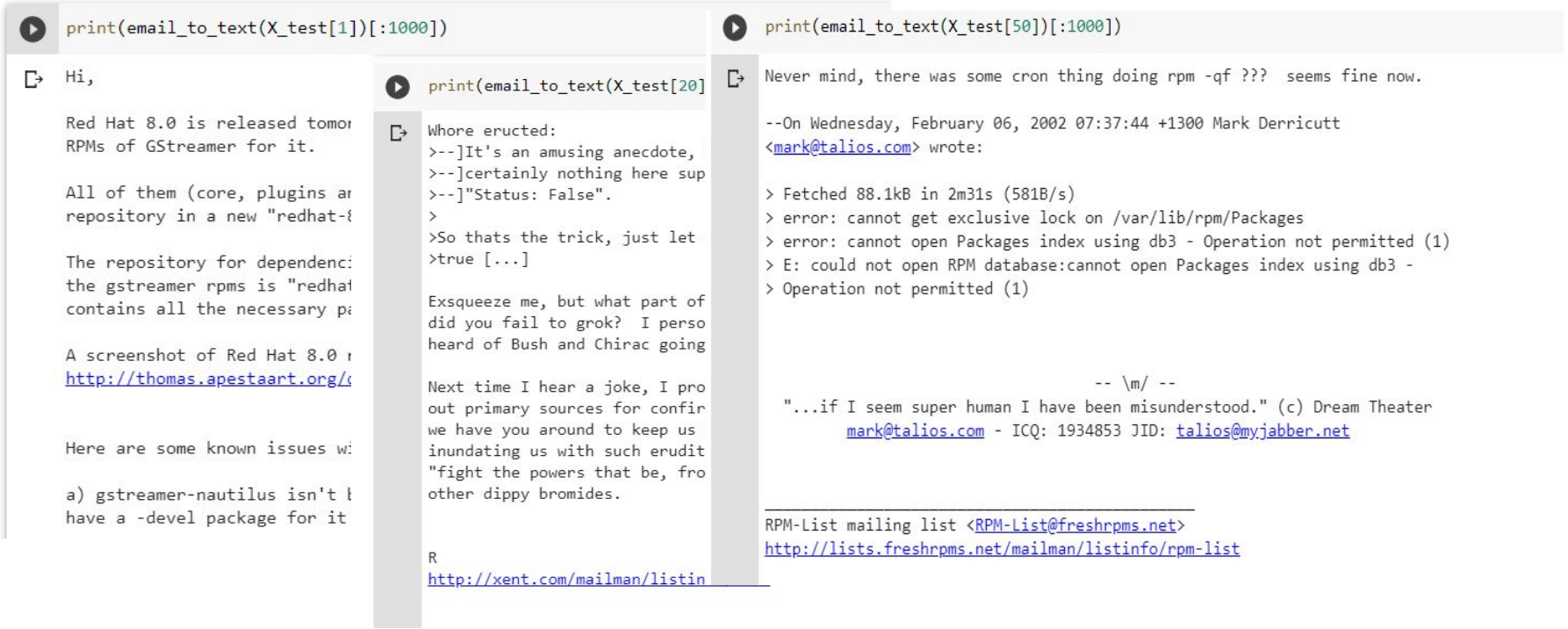
```
Mean cross validation score 0.9333333333333332
```

Hyperparameters for the model

```
dtc.get_params()
```

```
{'ccp_alpha': 0.0,
 'class_weight': None,
 'criterion': 'gini',
 'max_depth': 5,
 'max_features': None,
 'max_leaf_nodes': 5,
 'min_impurity_decrease': 0.0,
 'min_impurity_split': None,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'presort': 'deprecated',
 'random_state': None,
 'splitter': 'best'}
```

Some emails for testing



```
print(email_to_text(X_test[1])[:1000])
```

```
Hi,

Red Hat 8.0 is released tomor
RPMs of GStreamer for it.

All of them (core, plugins ar
repository in a new "redhat-8

The repository for dependenci
the gstreamer rpms is "redhat
contains all the necessary pa

A screenshot of Red Hat 8.0 r
http://thomas.apestaart.org/

Here are some known issues wi

a) gstreamer-nautilus isn't b
have a -devel package for it
```

```
print(email_to_text(X_test[20]
```

```
Whore eructed:
>--]It's an amusing anecdote,
>--]certainly nothing here sup
>--]"Status: False".
>
>So thats the trick, just let
>true [...]

Exsqueeze me, but what part of
did you fail to grok?  I perso
heard of Bush and Chirac going

Next time I hear a joke, I pro
out primary sources for confir
we have you around to keep us
inundating us with such erudit
"fight the powers that be, fro
other dippy bromides.


R
http://xent.com/mailman/listin
```

```
print(email_to_text(X_test[50])[:1000])
```

```
Never mind, there was some cron thing doing rpm -qf ???  seems fine now.

--On Wednesday, February 06, 2002 07:37:44 +1300 Mark Derricutt
<mark@talios.com> wrote:

> Fetched 88.1kB in 2m31s (581B/s)
> error: cannot get exclusive lock on /var/lib/rpm/Packages
> error: cannot open Packages index using db3 - Operation not permitted (1)
> E: could not open RPM database:cannot open Packages index using db3 -
> Operation not permitted (1)



                                        -- \m/ --
  "...if I seem super human I have been misunderstood." (c) Dream Theater
        mark@talios.com - ICQ: 1934853 JID: talios@myjabber.net


_____
RPM-List mailing list <RPM-List@freshrpms.net>
http://lists.freshrpms.net/mailman/listinfo/rpm-list
```

Decoding the word embeddings for better understanding

We need to find the values for X[116], X[135], X[27], X[78]

We do it using the word to vector transform in the pipeline

```
[153] preprocess_pipeline["wordcount_to_vector"].most_common_[116][0]

'e'
```

```
[166] preprocess_pipeline["wordcount_to_vector"].most_common_[135][0]

'these'
```

```
[167] preprocess_pipeline["wordcount_to_vector"].most_common_[27][0]

'list'
```

```
preprocess_pipeline["wordcount_to_vector"].most_common_[78][0]

'm'
```

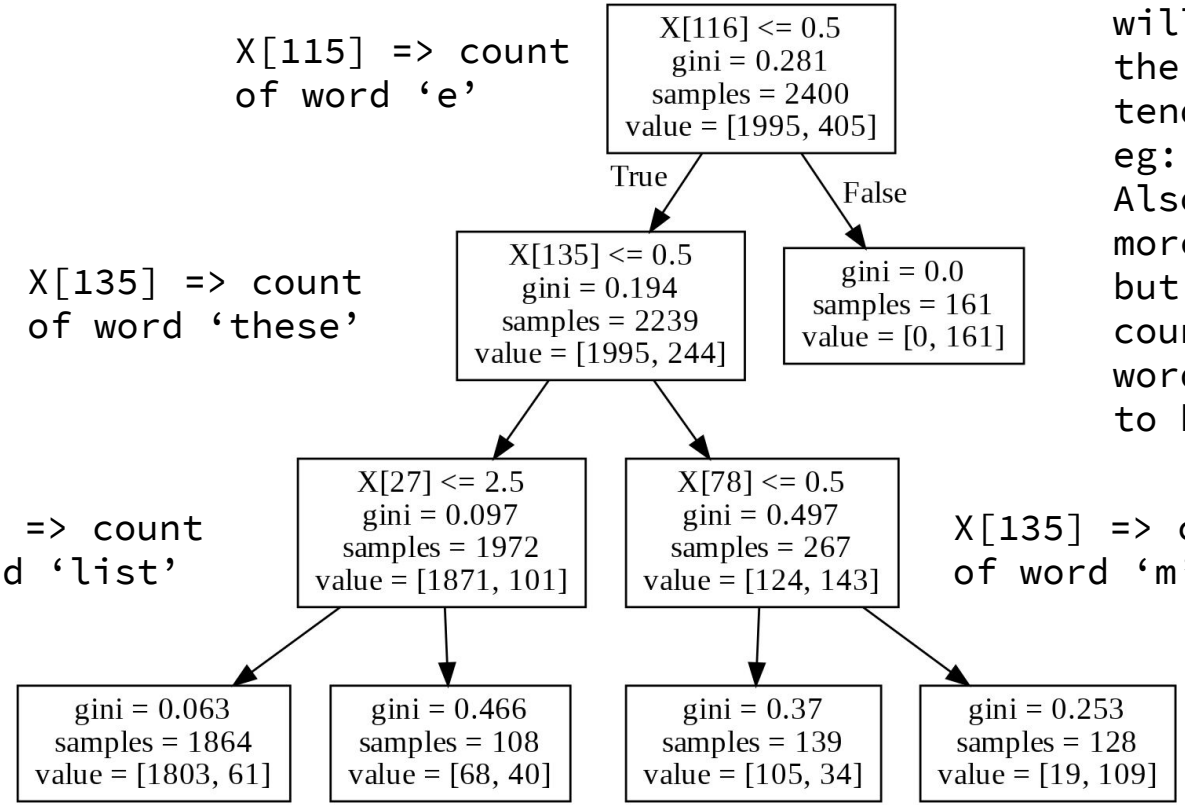
Decision Tree Generated By the Model

X[115] => count of word ‘e’

X[135] => count of word ‘these’

X[135] => count of word ‘list’

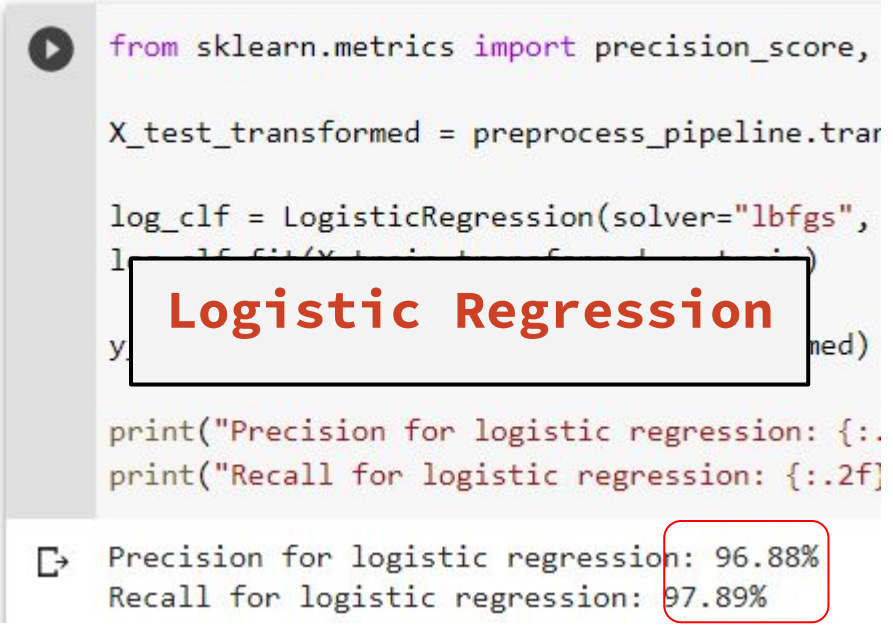
X[135] => count of word ‘m’

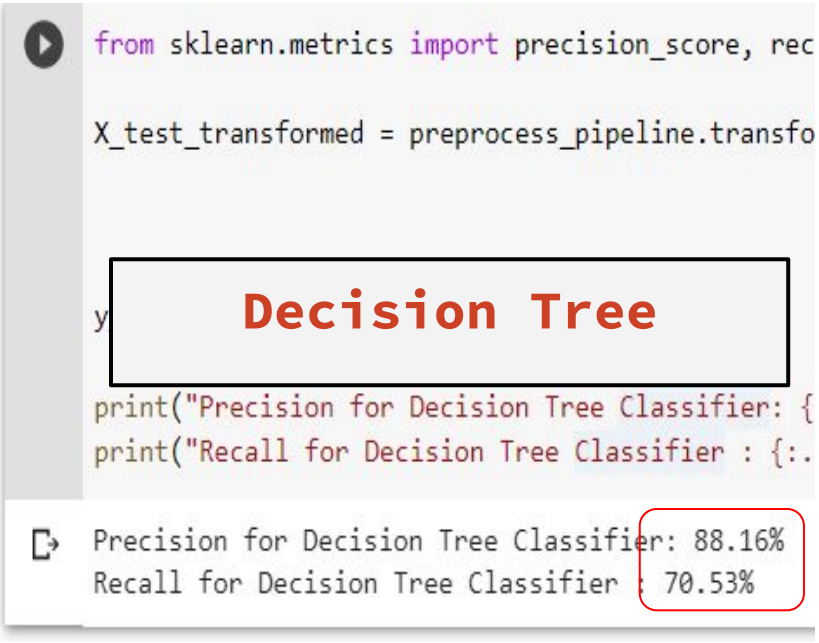


It seems from the model that mails will more count of the words with e tend to be spams eg: Fr**ee** ….
Also mails with more count of e but with higher count of these and words with m tend to be hams

Comparison with Logistic Regression on Test Set

Logistic Regression Performs Better on the test set. However we also see that even though Decision Tree performs poorly in comparison to Log Reg, it is more explainable





Example Predictions by the model

Actual: Ham
Predicted: Ham

```
y_pred[1], y_test[0]
(0, 0)
```

```
print(email_to_text(X_test[1])[:1
```

```
Hi,

Red Hat 8.0 is released tomorrow (monday).  I took some time out to make
RPMs of GStreamer for it.

All of them (core, plugins and player) have been uploaded to the apt
repository in a new "redhat-80-i386" directory.

The repository for dependencies is again called "deps", and the one for
the gstreamer rpms is "redhat", because this time around the base distro
contains all the necessary packages.

A screenshot of Red Hat 8.0 running gst-player is up at
http://thomas.apestaart.org/download/screenshots/redhat-80-gst-player.png


Here are some known issues with the resulting rpms :

a) gstreamer-nautilus isn't built, the package got renamed and I don't
have a -devel package for it yet
```

Actual: Spam
Predicted: Spam

```
[199] y_pred[4], y_test[4]
(1, 1)
```

```
print(email_to_text(X_test[4])[:1000])
```

```
A POWERHOUSE GIFTING PROGRAM You Don't Want To Miss!

  GET IN WITH THE FOUNDERS!
The MAJOR PLAYERS are on This ONE
For ONCE be where the PlayerS are
This is YOUR Private Invitation

EXPERTS ARE CALLING THIS THE FASTEST WAY
TO HUGE CASH FLOW EVER CONCEIVED
Leverage $1,000 into $50,000 Over and Over Again

THE QUESTION HERE IS:
YOU EITHER WANT TO BE WEALTHY
OR YOU DON'T!!!
WHICH ONE ARE YOU?
I am tossing you a financial lifeline and for your sake I
Hope you GRAB onto it and hold on tight For the Ride of youR life!

Testimonials

Hear what average people are doing their first few days:
�We've received 8,000 in 1 day and we are doing that over and over again!'
 �I'm a single mother in FL and I've received 12,000 in the last 4 days.�
�I was not sure about this when I sent off my $1,000 pledge, but I got bac
�I didn't have the money, so I found myself a partner to work this with. W
```

Confusion Matrix for Decision Tree Model