Daniel Swieca
October 20th, 2025

Ham/Spam email detection

Introduction

The purpose of this project is to use our data science and machine learning knowledge accrued over the last several months to train a Machine Learning model that can categorize emails as either Spam or Not Spam (from now on, Not Spam will be referred to as Ham). At the end of this project, I had built a simple applet in which the user can input the contents of an email, and the applet processes the email using a trained machine learning model to categorize it as Spam or Ham.

Data Selection

Before a machine learning model can be trained, I was required to find an appropriate dataset with which to train it. I had chosen the Enron Spam Dataset, as it was a large and diverse enough dataset with over 33,000 entries, with a close balance of Ham and Spam email examples (16,614 spam entries and 16,493 ham entries).

Using an evenly balanced dataset was beneficial, reducing the amount of cleanup necessary during the data preparation phase, and leading to a more reliable final model when the training was done.

Data Cleanup

The structure of the dataset initially was:

Message	**Spam/Ham**
The content of the email, stored as a text value.	Whether this email is ham or spam, stored as a text value in the form of the words: Ham Spam

The data required cleanup before it could be used in a machine learning model. First, any rows with empty fields were dropped so as to not cloud the algorithm. Next, machine learning modules are incapable of reading text by default, and so all text fields had to be converted into a numeric format.

Initially, the Spam/Ham column was easily mapped to a boolean representation. All Ham values were converted to a 0, and all spam values were converted to a 1.

Next, the message field needed to be cleaned up. The first step is to remove any unnecessary or excessive information found in the message content. First, remove any words that aren't alpha. Meaning, any sections of the email that are either strictly conjugation, or numeric values.

Secondly, remove any "stop words" found in the email content. Stop words are common words (e.g., "the", "a", "is", "and") that are often removed during text preprocessing in Natural Language Processing (NLP) because they typically carry little semantic meaning and unnecessarily complicate our analysis.

At the end of this cleanup, the table would have transformed from this:

<table>
<tr><th>index</th><th>Message</th><th>Spam/Ham</th></tr>
<tr><td>1</td><td>gary , production from the high island larger block a - 1 # 2 commenced on saturday at 2 : 00 p . m . at about 6 , 500 gross . carlos expects between 9 , 500 and 10 , 000 gross for tomorrow . vastar owns 68 % of the gross production . george x 3 - 6992 - forwarded by george weissman / hou / ect on 12 / 13 / 99 10 : 16 am - daren j farmer 12 / 10 / 99 10 : 38 am to : carlos j rodriguez / hou / ect @ ect cc : george weissman / hou / ect @ ect , melissa graves / hou / ect @ ect subject : vastar resources , inc . carlos , please call linda and get everything set up . i ' m going to estimate 4 , 500 coming up tomorrow , with a 2 , 000 increase each following day based on my conversations with bill fischer at bmar . d . - forwarded by daren j farmer / hou / ect on 12 / 10 / 99 10 : 34 am - enron north america corp . from : george weissman 12 / 10 / 99 10 : 00 am to : daren j farmer / hou / ect @ ect cc : gary bryan / hou / ect @ ect , melissa graves / hou / ect @ ect subject : vastar resources , inc . darren , the attached appears to be a nomination from vastar resources , inc . for the high island larger block a - 1 # 2 (previously , erroneously referred to as the # 1 well) . vastar now expects the well to commence production sometime tomorrow . i told linda harris that we ' d get her a telephone number in gas control so she can provide notification of the turn - on tomorrow . linda ' s numbers , for the record , are 281 . 584 . 3359 voice and 713 . 312 . 1689 fax . would you please see that someone contacts linda and advises her how to submit future nominations via e - mail , fax or voice ? thanks . george x 3 - 6992 - forwarded by george weissman / hou / ect on 12 / 10 / 99 09 : 44 am - " linda harris " on 12 / 10 / 99 09 : 38 : 43 am to : george weissman / hou / ect @ ect cc : subject : hi a - 1 # 2 effective 12 - 11 - 99 | - - - - - - - - + - - - - - - - - - - + - - - - - - - - - - - | | | | | | mscf / d | min ftp | time | | | | | | - - - - - - - - + - - - - - - - - - - + - - - - - - - - - - - | | | | | | 4 , 500 | 9 , 925 | 24 hours | | | | | | - - - - - - - - + - - - - - - - - - - + - - - - - - - - - - - | | | | | | 6 , 000 | 9 , 908 | 24 hours | | | | | | - - - - - - - - + - - - - - - - - - - + - - - - - - - - - - - | | | | | | 8 , 000 | 9 , 878 | 24 hours | | | | | | - - - - - - - - + - - - - - - - - - - + - - - - - - - - - - - | | | | | | 10 , 000 | 9 , 840 | 24 hours | | | | | | - - - - - - - - + - - - - - - - - - - + - - - - - - - - - - - | | | | | | 12 , 000 | 9 , 793 | 24 hours | | | | | | - - - - - - - - + - - - - - - - - - - + - - - - - - - - - - - | | | | | | 14 , 000 | 9 , 738 | 24 hours | | | | | | - - - - - - - - + - - - - - - - - - - + - - - - - - - - - - - | | | | | | 16 , 000 | 9 , 674 | 24 hours | | | | | | - - - - - - - - + - - - - - - - - - - + - - - - - - - - - - - | | | | | | 18 , 000 | 9 , 602 | 24 hours | | | | | | - - - - - - - - + - - - - - - - - - - + - - - - - - - - - - - | | | | | | 20 , 000 | 9 , 521 | 24 hours | | | | | | - - - - - - - - + - - - - - - - - - - + - - - - - - - - - - - | | | | | | 22 , 000 | 9 , 431 | 24 hours | | | | | | - - - - - - - - + - - - - - - - - - - + - - - - - - - - - - - | | | | | | 24 , 000 | 9 , 332 | 24 hours | | | | | | - - - - - - - - + - - - - - - - - - - + - - - - - - - - - - - | | | | | | 26 , 000 | 9 , 224 | 24 hours | | | | | | - - - - - - - - + - - - - - - - - - - + - - - - - - - - - - - | | | | | | 28 , 000 | 9 , 108 | 24 hours | | | | | | - - - - - - - - + - - - - - - - - - - + - - - - - - - - - - - | | | | | | 30 , 000 | 8 , 982 | 24 hours | | | | | | - - - - - - - - + - - - - - - - - - - + - - - - - - - - - - - | | | | | | 32 , 000 | 8 , 847 | 24 hours | | | | | | - - - - - - - - + - - - - - - - - - - + - - - - - - - - - - - | | | | | | 34 , 000 | 8 , 703 | 24 hours | | | | | | - - - - - - - - + - - - - - - - - - - + - - - - - - - - - - - | | | | | | 36 , 000 | 8 , 549 | 24 hours | | | | | | - - - - - - - - + - - - - - - - - - - + - - - - - - - - - - - |</td><td>ham</td></tr>
<tr><td>2</td><td>- calpine daily gas nomination 1 . doc</td><td>ham</td></tr>
<tr><td>3</td><td>fyi - see note below - already done . stella - forwarded by stella l morris / hou / ect on 12 / 14 / 99 10 : 18 am - from : sherlyn schumack on 12 / 14 / 99 10 : 06 am to : stella l morris / hou / ect @ ect cc : howard b camp / hou / ect @ ect subject : re : issue stella , this has already been taken care of . you did this for me yesterday . thanks . howard b camp 12 / 14 / 99 09 : 10 am to : stella l morris / hou / ect @ ect cc : sherlyn schumack / hou / ect @ ect , howard b camp / hou / ect @ ect , stacey neuweiler / hou / ect @ ect , daren j farmer / hou / ect @ ect subject : issue stella , can you work with stacey or daren to resolve hc - forwarded by howard b camp / hou / ect on 12 / 14 / 99 09 : 08 am - from : sherlyn schumack 12 / 13 / 99 01 : 14 pm to : howard b camp / hou / ect @ ect cc : subject : issue i have to create accounting arrangement for purchase from unocal energy at meter 986782 . deal not tracked for 5 / 99 . volume on deal 114427 expired 4 / 99 .</td><td>ham</td></tr>
<tr><td>4</td><td>fyi . - forwarded by lauri a allen / hou / ect on 12 / 14 / 99 12 : 17 pm - kimberly vaughn 12 / 10 / 99 02 : 54 pm to : lauri a allen / hou / ect @ ect cc : mary m smith / hou / ect @ ect subject : meter 7268 nov allocation lauri . . i have put this on strangas gas until i can get a contract from daren . - forwarded by kimberly vaughn / hou / ect on 12 / 10 / 99 01 : 52 pm - lauri a allen 12 / 09 / 99 01 : 20 pm to : kimberly vaughn / hou / ect @ ect , anita luong / hou / ect @ ect cc : howard b camp / hou / ect @ ect , mary m smith / hou / ect @ ect subject : meter 7268 nov allocation kim / anita - a volume of 7247 mm shows to have been allocated to the reliant 201 contract for november . there was no nomination for reliant at this point in november and , therefore , there should be no volume allocated to their contract . please make sure these volumes are moved off the reliant contract prior to november close . thanks .</td><td>ham</td></tr>
<tr><td>5</td><td>jackie , since the inlet to 3 river plant is shut in on 10 / 19 / 99 (the last day of flow) : at what meter is the mcmullen gas being diverted to ? at what meter is hpl buying the residue gas ? (this is the gas from teco , vastar , vintage , tejones , and swift) i still see active deals at meter 3405 in path manager for teco , vastar , vintage , tejones , and swift i also see gas scheduled in pops at meter 3404 and 3405 . please advice . we need to resolve this as soon as possible so settlement can send out payments . thanks</td><td>ham</td></tr>
</table>

To this:

index	Message	Spam_Bool
1	gari product high island larger block commenc saturday p gross carlo expect gross tomorrow vastar own gross product georg x forward georg weissman hou ect daren j farmer carlo j rodriguez hou ect ect cc georg weissman hou ect ect melissa grave hou ect ect subject vastar resourc inc carlo pleas call linda get everyth set go estim come tomorrow increas follow day base convers bill fischer bmar forward daren j farmer hou ect enron north america corp georg weissman daren j farmer hou ect ect cc gari bryan hou ect ect melissa grave hou ect ect subject vastar resourc inc darren attach appear nomin vastar resourc inc high island larger block previous erron refer well vastar expect well commenc product sometim tomorrow told linda harri get telephon number ga control provid notif turn tomorrow linda number record voic fax would pleas see someon contact linda advis submit futur nomin via e mail fax voic thank georg x forward georg weissman hou ect linda harri georg weissman hou ect ect cc subject hi effect mscf min ftp time hour hour hour hour hour hour hour hour hour hour hour hour hour hour hour hour hour	0
2	calpin daili ga nomin doc	0
3	fyi see note alreadi done stella forward stella l morri hou ect sherlyn schumack stella l morri hou ect ect cc howard b camp hou ect ect subject issu stella alreadi taken care yesterday thank howard b camp stella l morri hou ect ect cc sherlyn schumack hou ect ect howard b camp hou ect ect stacey neuweil hou ect ect daren j farmer hou ect ect subject issu stella work stacey daren resolv hc forward howard b camp hou ect sherlyn schumack pm howard b camp hou ect ect cc subject issu creat account arrang purchas unoc energi meter deal track volum deal expir	0
4	fyi forward lauri allen hou ect pm kimberli vaughn pm lauri allen hou ect ect cc mari smith hou ect ect subject meter nov alloc lauri put stranga ga get contract daren forward kimberli vaughn hou ect pm lauri allen pm kimberli vaughn hou ect ect anita luong hou ect ect cc howard b camp hou ect ect mari smith hou ect ect subject meter nov alloc kim anita volum mm show alloc reliant contract novemb nomin reliant point novemb therefor volum alloc contract pleas make sure volum move reliant contract prior novemb close thank	0
5	jacki sinc inlet river plant shut last day flow meter mcmullen ga divert meter hpl buy residu ga ga teco vastar vintag tejon swift still see activ deal meter path manag teco vastar vintag tejon swift also see ga schedul pop meter pleas advic need resolv soon possibl settlement send payment thank	0

The alpha, stemmed messages only have the key information and lack any of the noise in the form of punctuation and formatting found in the original dataset. This simplifies them and decreases the number of features necessary for our machine learning algorithm to analyse.

The final step in the data cleaning step, is to convert these alpha values into numerical tokens so that an ML algorithm can read them properly. This was done using Sci-Kit Learn's CountVectorizer. After all data was cleaned, this was the final structure of the dataset table.

Message	**Spam_Bool**
The content of the email, stored as a numerical token value.	Whether this email is ham or spam, stored as a numerical value in the form of the numbers: 0 1

Model Training

After cleaning all the data, a machine learning model needed to be trained. After trying three separate models, the Logistic Regression algorithm proved to be the most accurate, with the fewest false positive values and the fewest false negative values.

The final trained Logistic Regression model had the following evaluation:

Confusion Matrix:

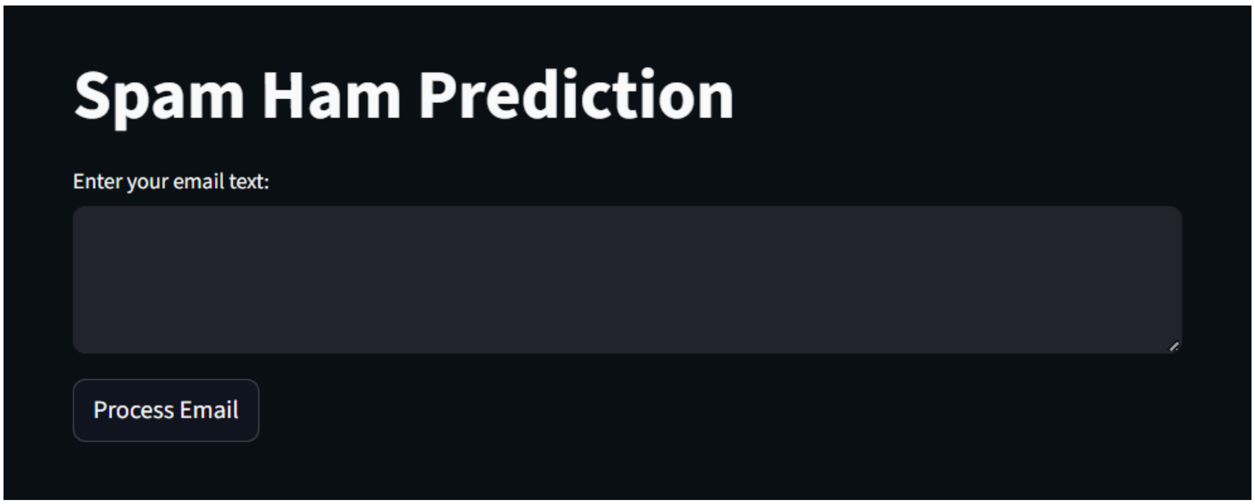
True Positives: 4876	**False Positives:** 87
False Negatives: 46	**True Negatives**: 4924

Accuracy: 0.9866
Precision: 0.9906
Recall: 0.9825
f1-score: 0.9865

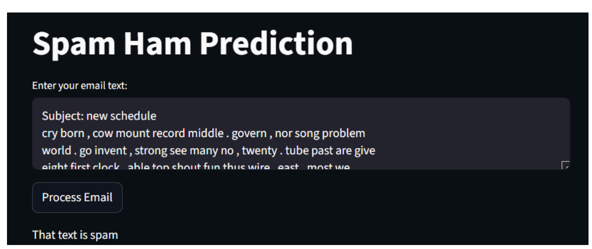
This was a very satisfactory result, with a near-perfect Precision of over 99%. Satisfied with the model performance all that was left was to implement it inside an app.

Streamlit Application

The model was saved to a pickle file, and integrated into a simple streamlit applet.



The user can input any email text content into the textbox, and it'll categorize it into Spam or Ham using the previously-trained Machine Learning model. The sole complication I had not foreseen with the implementation into the streamlit applet, is that the same tokenizer used to initially tokenize the data had to be reused in the later applet. This was easily handled by saving the tokenizer to an external file in the transformation step, and loading it in the streamlit applet.



Reflection

This project was an interesting exercise for me as it felt like my first time having to really do the entire ML pipeline with minimal guidance. I had to personally do the whole flow of finding a dataset, cleaning it, training a model with it, and then implementing it into a presentable app prototype. This project left me feeling more prepared for future Machine Learning projects, and steps that took longer this time around (most notably, data cleanup) I think will be much faster moving forward as the sense of confidence has seriously increased.