Can the Spam: Detecting Junk Emails Using Machine Learning

By Renae Alailima, Saharsh Bhargava, Megan Nguyen, Somrat Sen

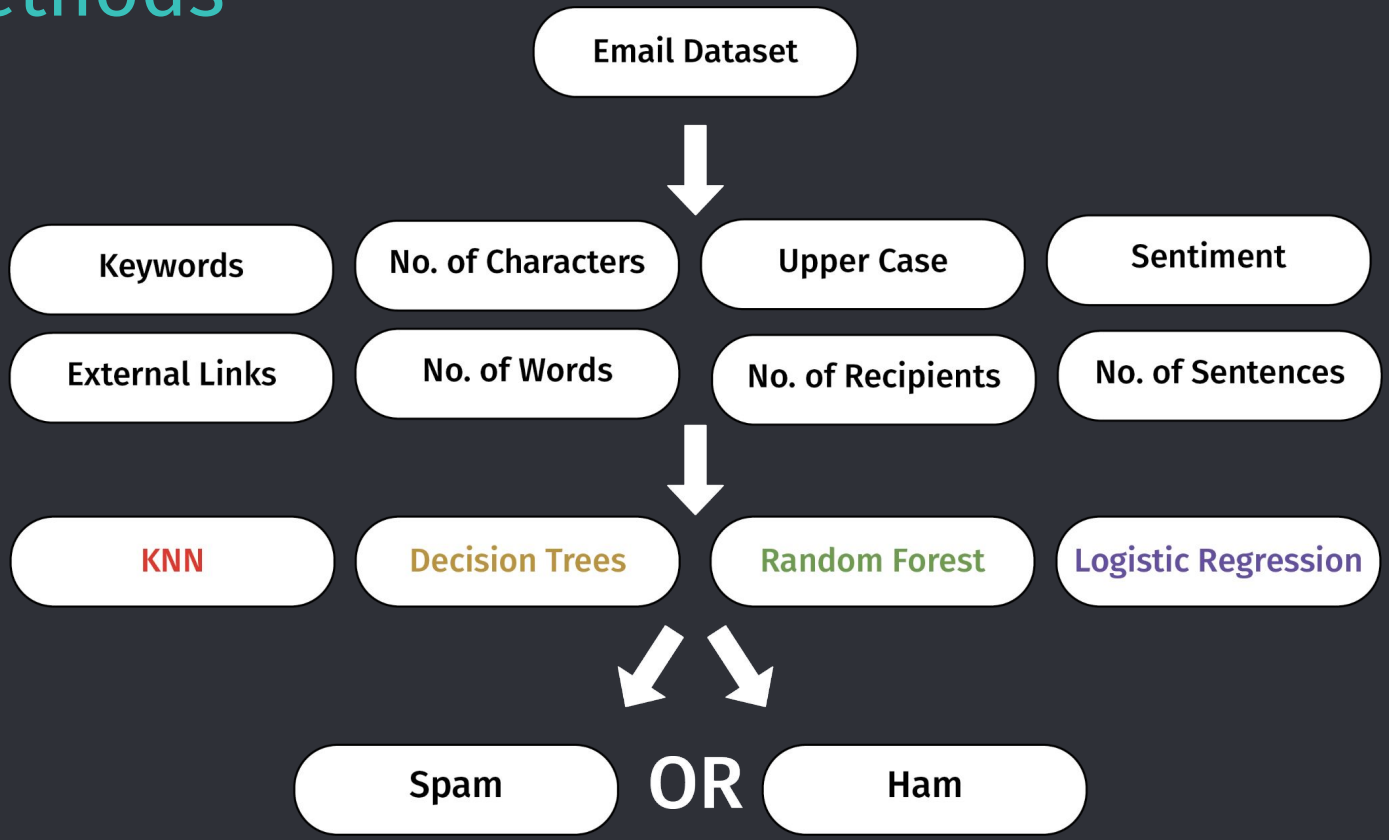
Purpose

Materials

SourceFile	Subject	Recipients	Sender	RawEmailBody	PlainTextEmailBody	label
enron_mail_20150507.tar/maildir/arnold-j/_sent...	Re: good morning	jenwhite7@zdnetonebox.com	john.arnold@enron.com	my polling location is the knights of columbus...	my polling location is the knights of columbus...	ham
enron_mail_20150507.tar/maildir/symes-k/sent/849.	Re: 3/13 Checkout	evelyn.metoyer@enron.com	kate.symes@enron.com	We've changed this to a BOM deal (term 3/15 to...	We've changed this to a BOM deal (term 3/15 to...	ham
enron_mail_20150507.tar/maildir/gilbertsmith-d...	total charges for installation	doug.gilbert-smith@enron.com	chris.clark@enron.com	Hey Doug - The total charges related to the eq...	Hey Doug - The total charges related to the eq...	ham
enron_mail_20150507.tar/maildir/kean-s/all_doc...	Energy Issues	elizabeth.linnell@enron.com, filuntz@aol.com, ...	miyung.buster@enron.com	Please see the attached articles:\n\n\n\n\n\n\...	Please see the attached articles:\n\n\n\n\n\n\...	ham
enron_mail_20150507.tar/maildir/ward-k/gas_cus...	Generation Development Term Sheet: REU's Clea...	tnichols@ci.redding.ca.us	laird.dyer@enron.com	Tim,\n\nFurther to our discussion this morning...	Tim,\n\nFurther to our discussion this morning...	ham

500,000+ Emails & 7,500+ Spam Emails

Methods

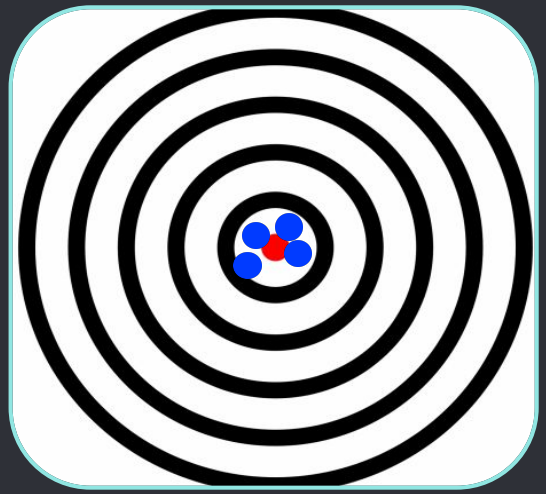

Email Dataset
Keywords
No. of Characters
Upper Case
Sentiment
External Links
No. of Words
No. of Recipients
No. of Sentences
KNN
Decision Trees
Random Forest
Logistic Regression
Spam
OR
Ham

Definitions

Spam → Malicious Emails

Ham → Regular Emails

Accuracy + Precision

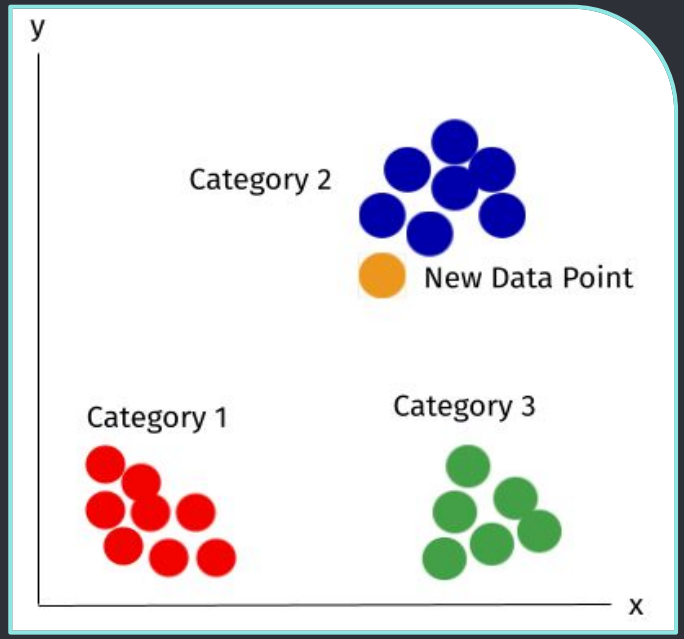


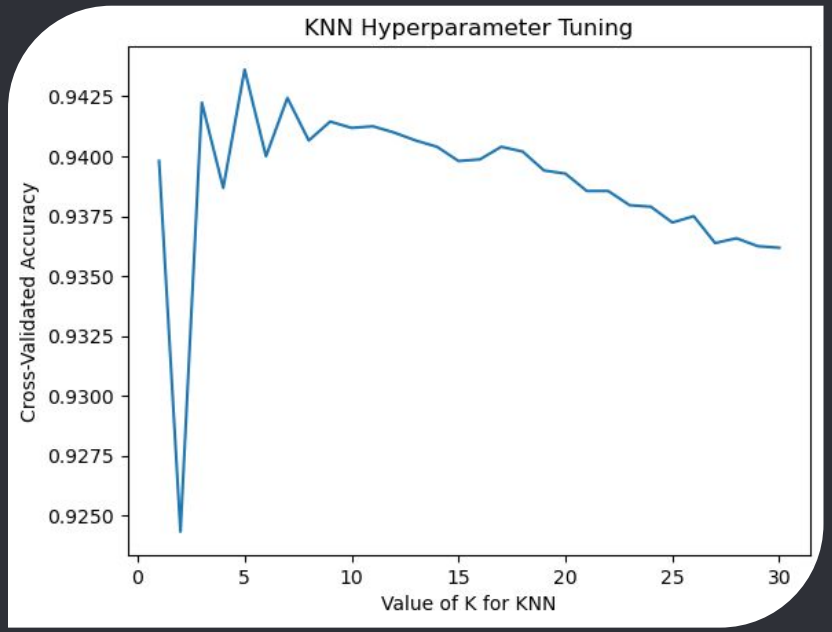
Precision Only

Neither

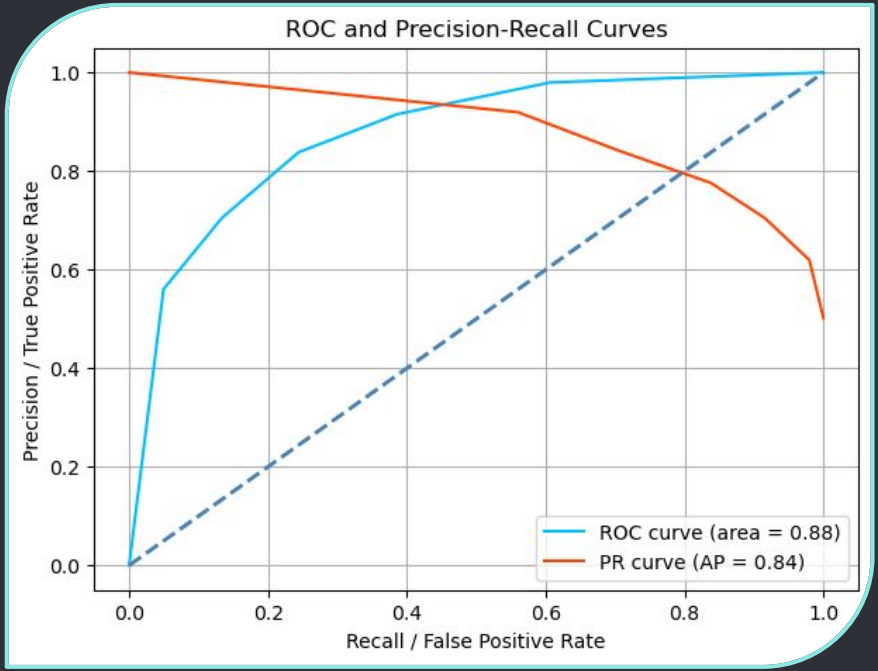
Accuracy Only

K-Nearest Neighbors (KNN)





K-Nearest Neighbors Results

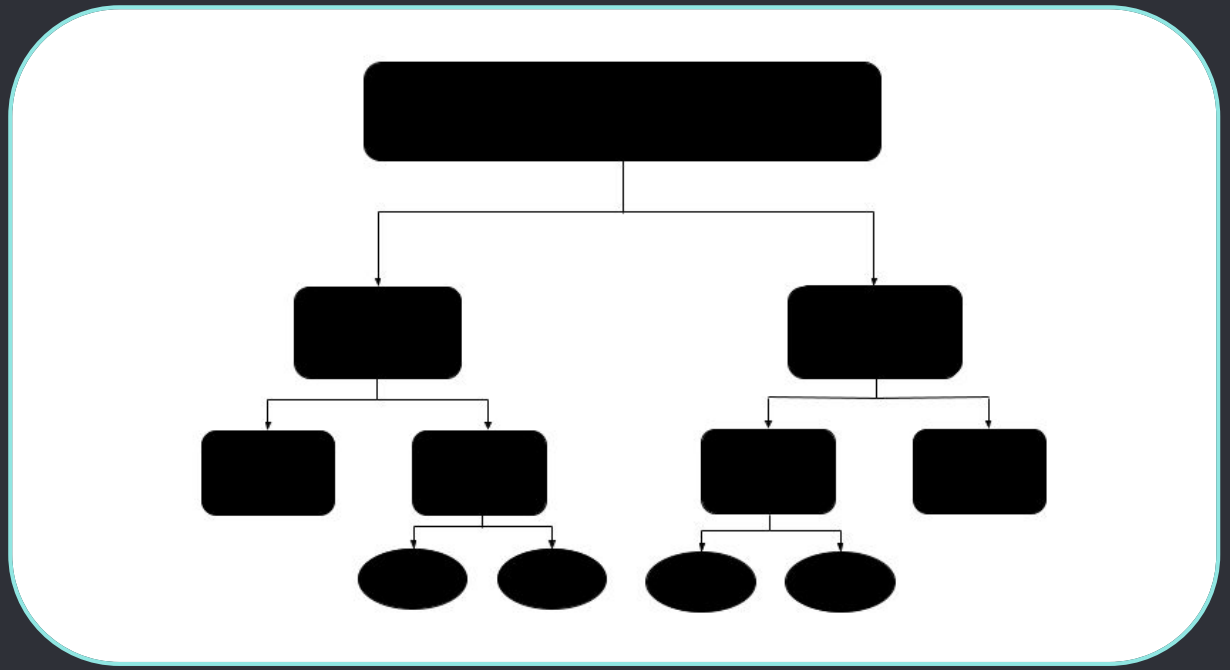


Accuracy: 80%

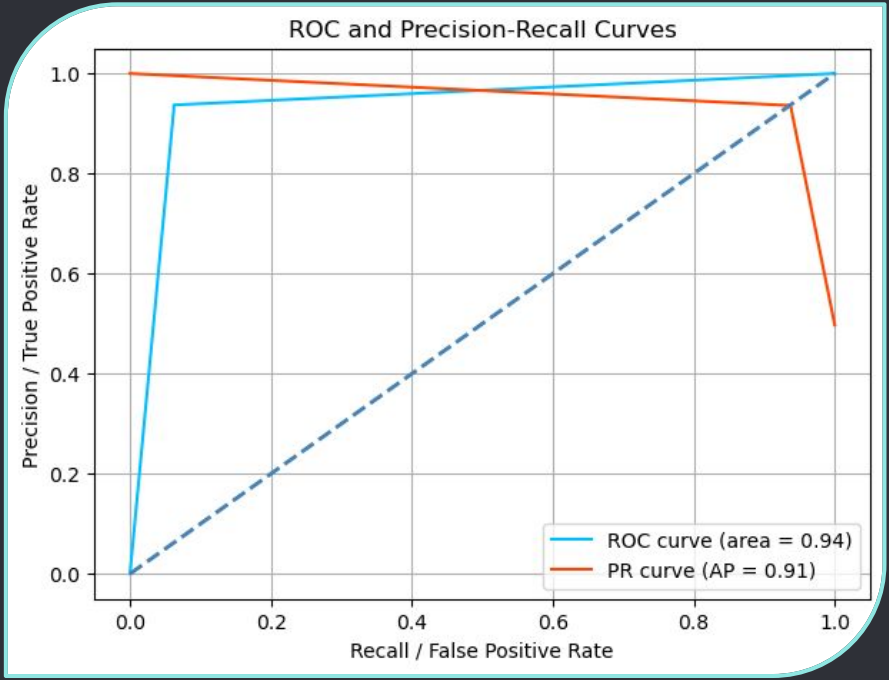
Precision: 83%

Recall: 83%

Decision Trees



Decision Trees Results

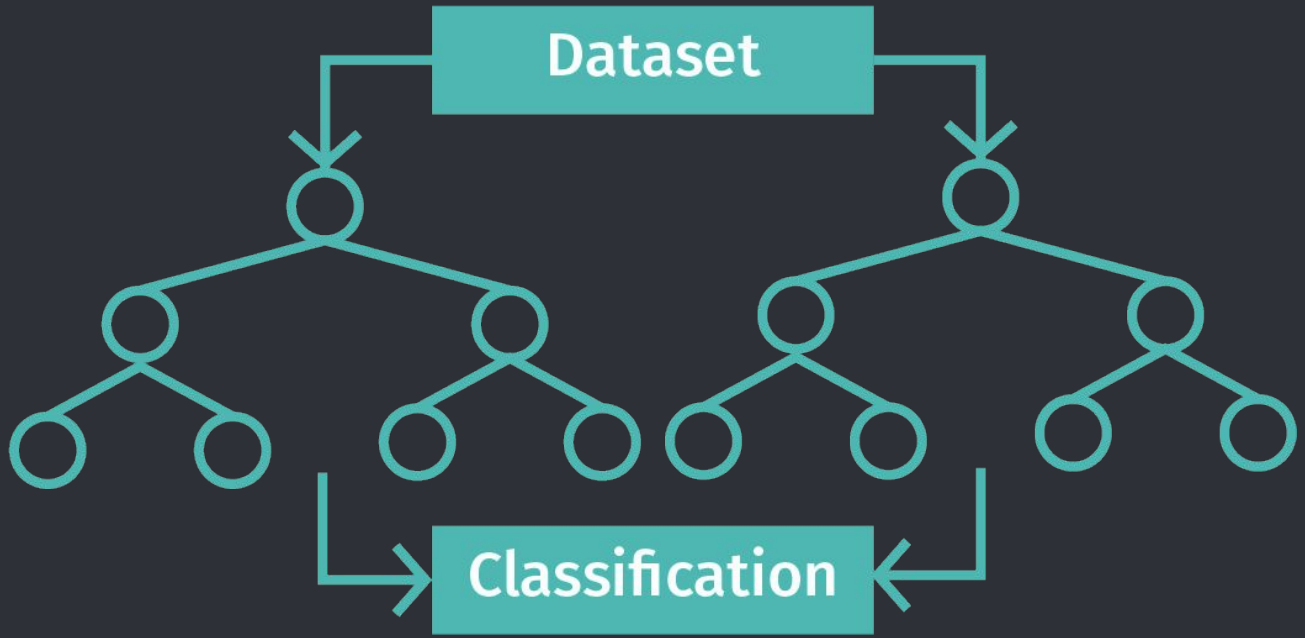


Accuracy: 94%

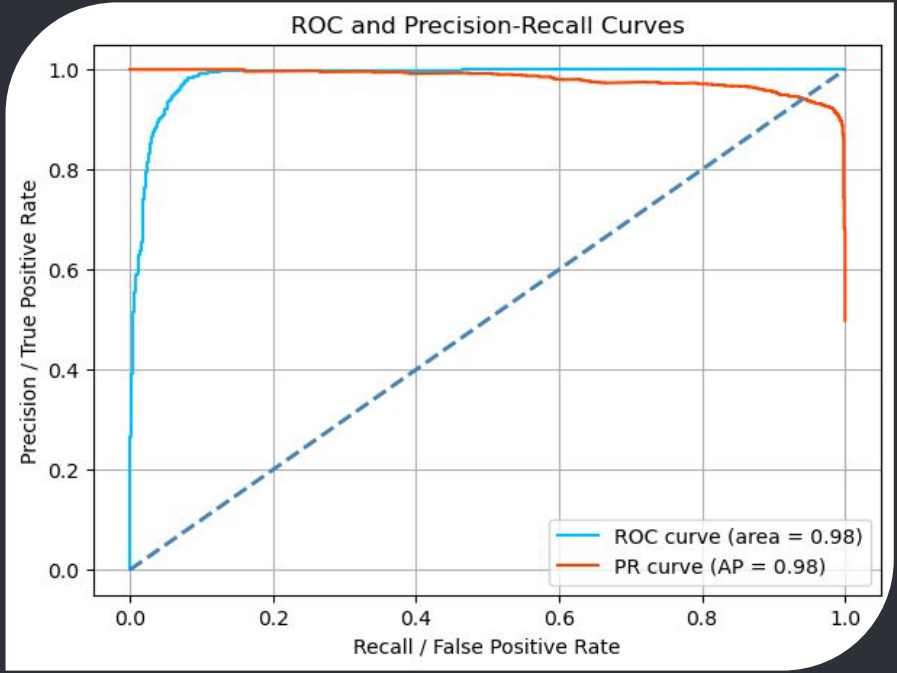
Precision: 94%

Recall: 94%

Random Forest



Random Forest Results

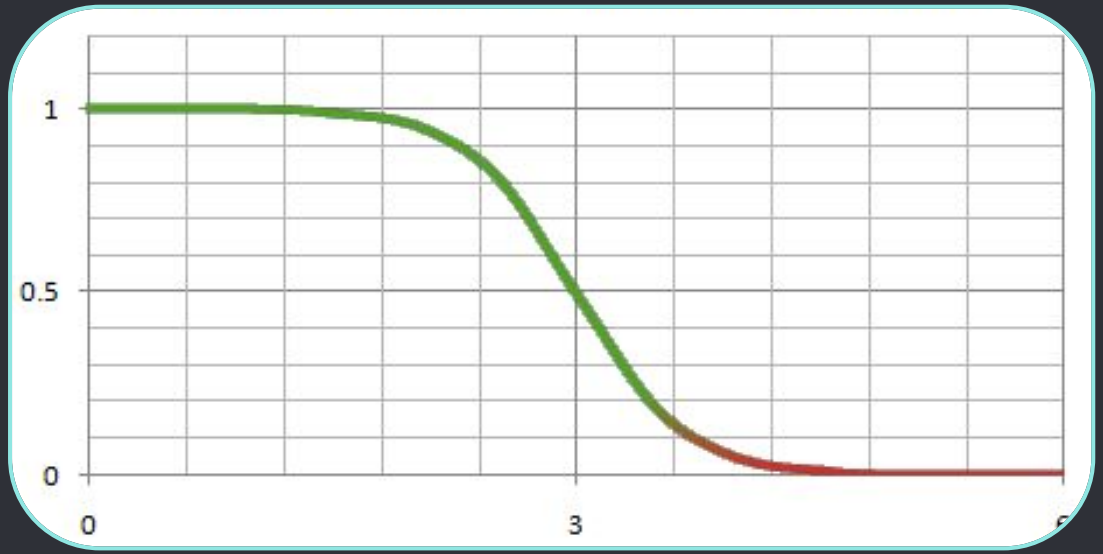


Accuracy: 96%

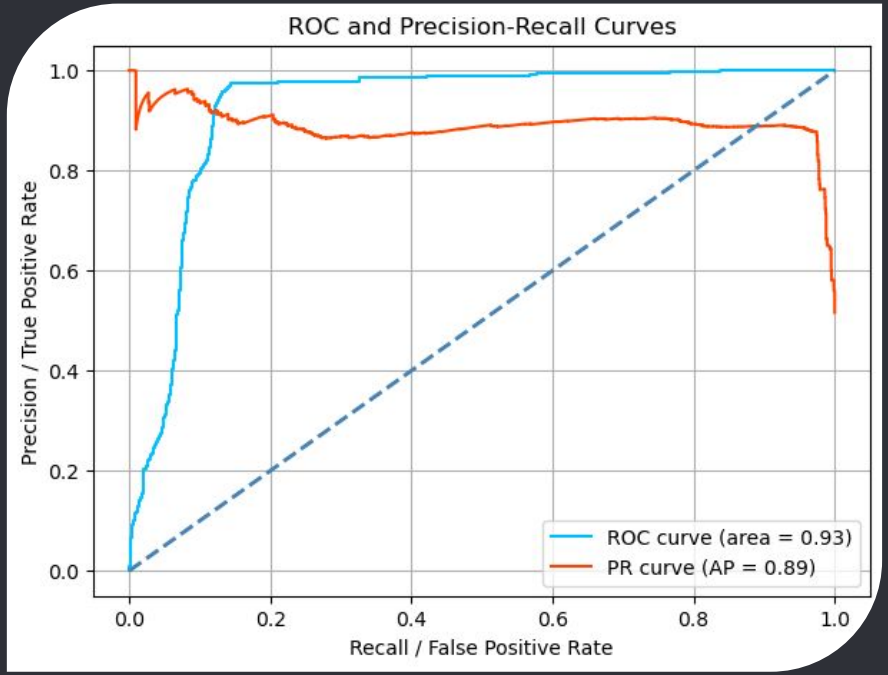
Precision: 98%

Recall: 94%

Logistic Regression



Logistic Regression Results



Accuracy: 92%

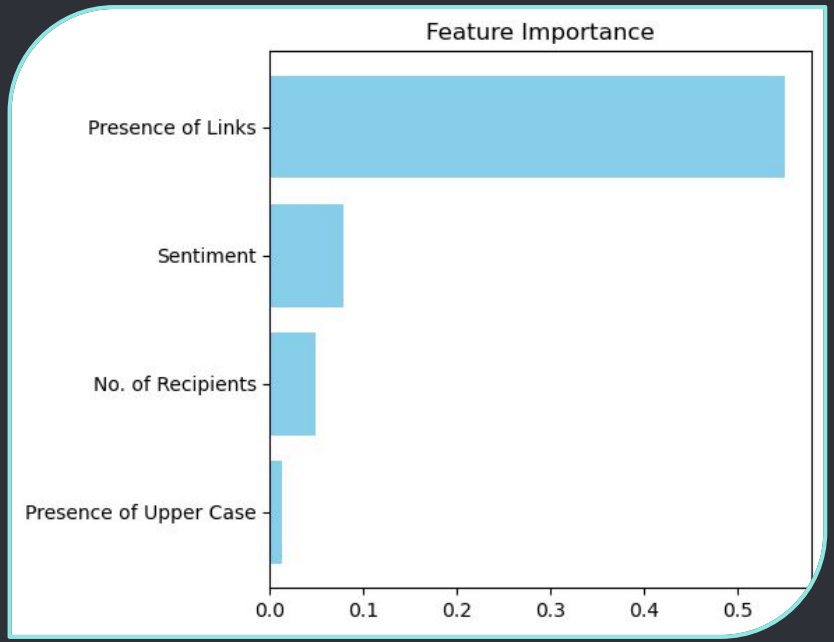
Precision: 97%

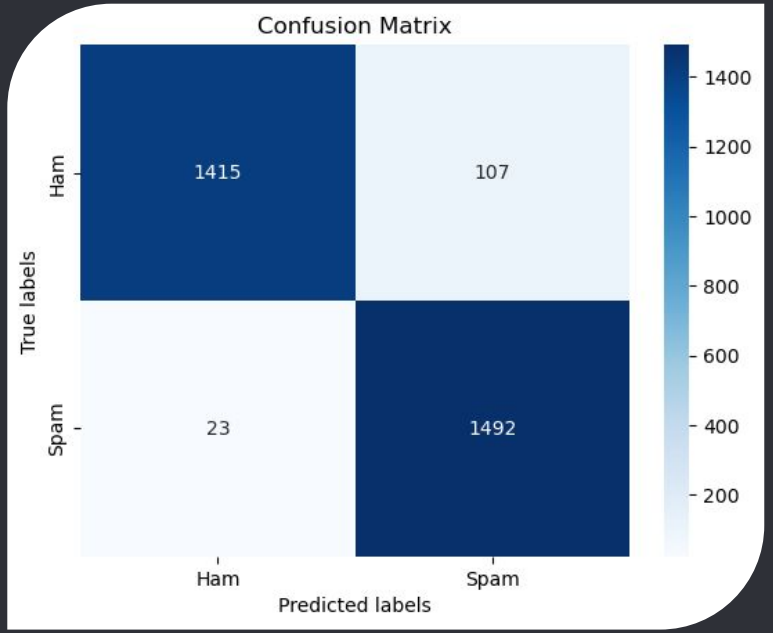
Recall: 88%

Combined Metrics

	KNN	Decision Trees	Random Forest	Logistic Regression
Accuracy	79.55%	94.34%	***95.79%***	91.90%
Precision	82.58%	94.33%	***98.07%***	96.69%
Recall	82.57%	***94.43%***	94.06%	88.17%
AUC	0.88	0.94	***0.98***	0.93
Ranking	4th	2nd	1st	3rd

Analysis


Feature Importance
Presence of Links
Sentiment
No. of Recipients
Presence of Upper Case
0.0
0.1
0.2
0.3
0.4
0.5


Confusion Matrix
True labels
Ham
Spam
1415
107
23
1492
Ham
Spam
Predicted labels
1400
1200
1000
800
600
400
200

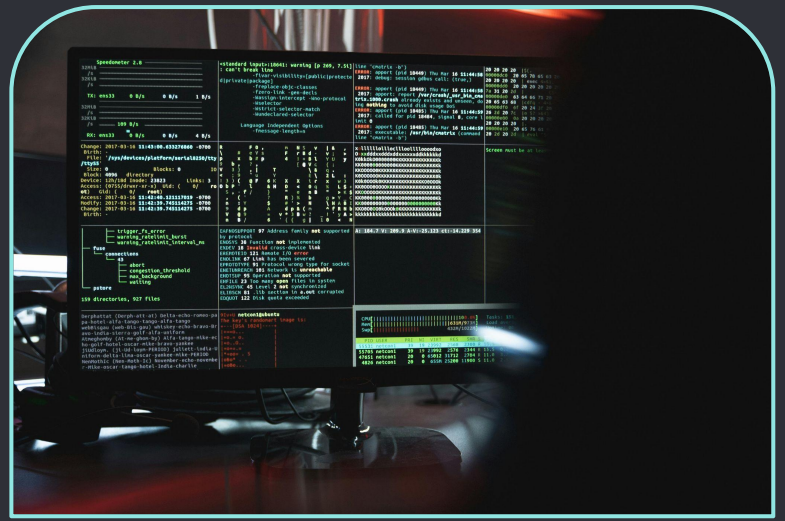
Drawbacks

Limitations

Computing Power

Biased Dataset

Machine Learning in the Future



Thank You!!

Head Mentor Meifan Chen, Mentors Bryan Arguello, Carol Chen, Brian Gaume, Tian Ma, Doug McGeehan, Anton Sumali, Alumni Assistant Peter Escamilla, Sandia National Laboratories, Oak Ridge Institute for Science and Education, The University of New Mexico, and the Defense Threat Reduction Agency's Joint Science and Technology Office for Chemical and Biological Defense.