Classification

Introduction

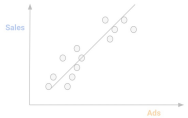
Prof. Dr. Jan Kirenz

There are different types of models

Use the patterns in my data to make predictions

Supervised learning

Predicting numerical values



Regression

Linear Regression, Decision Tree, Random Forest, Gradient Boosting Tree

Predicting categories

Conversion | No conversion

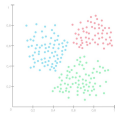
Classification

Logistic Regression, Decision Tree, Random Forest, Gradient Boosting Tree

Tell me what patterns exist in my data

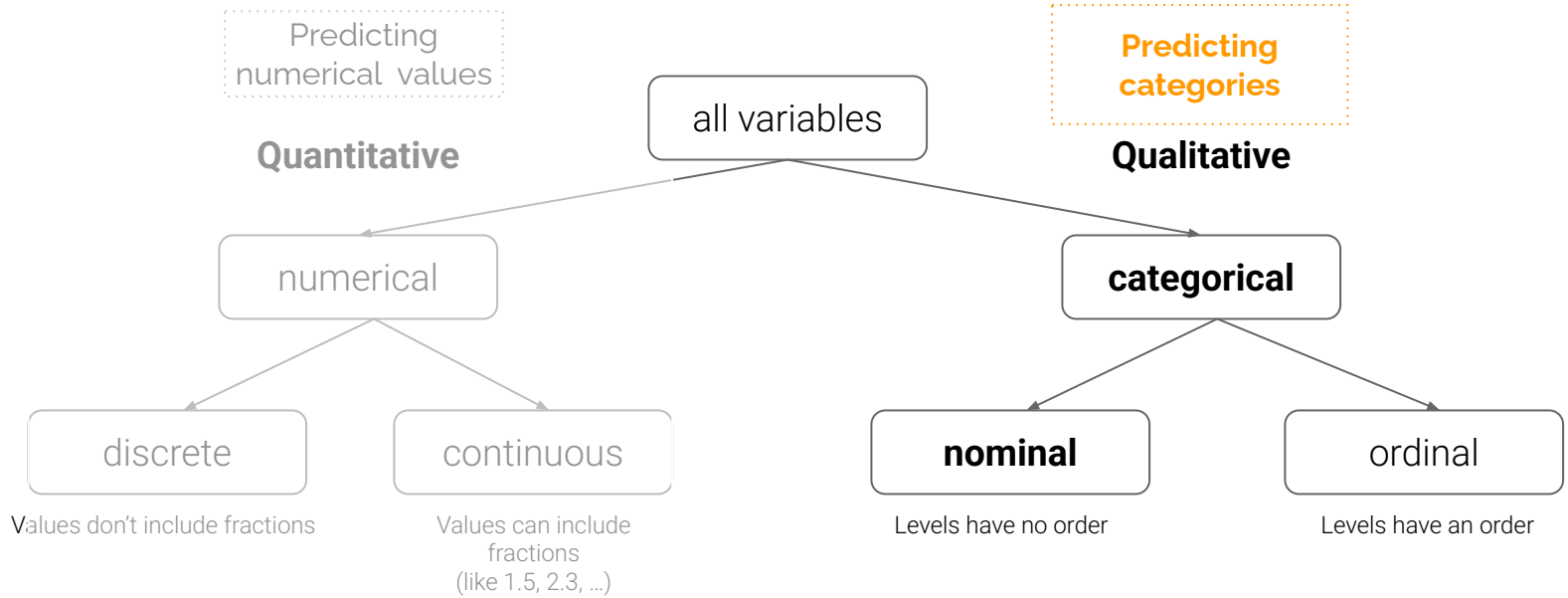
Unsupervised learning

Discovering structure



Clustering

Breakdown of **variables** into their **types**



Source: Çetinkaya-Rundel & Hardin (2021)

Predicting a category for some given input

Binary classification

Conversion

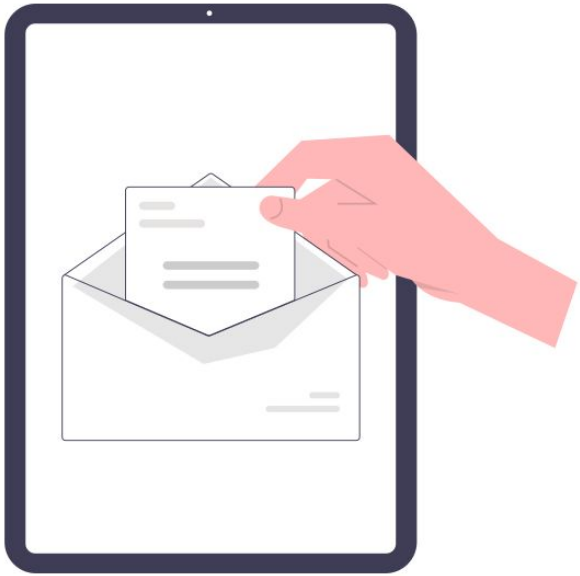
No conversion

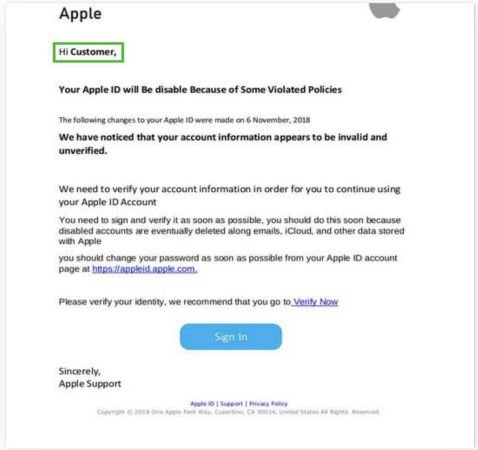
Multiclass classification (3 or more)

2 or more conversions

One conversion

No conversion





Apple

Hi Customer,

Your Apple ID will Be disable Because of Some Violated Policies

The following changes to your Apple ID were made on 6 November, 2018

We have noticed that your account information appears to be invalid and unverified.

We need to verify your account information in order for you to continue using your Apple ID Account

You need to sign and verify it as soon as possible, you should do this soon because disabled accounts are eventually deleted along emails, iCloud, and other data stored with Apple

you should change your password as soon as possible from your Apple ID account page at https://appleid.apple.com.

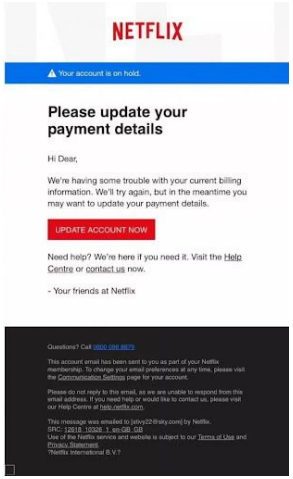
Please verify your identity, we recommend that you go to Verify Now

Sign In

Sincerely,
Apple Support

Apple ID | Support | Privacy Policy

Copyright © 2018 One Apple Park Way, Cupertino, CA 95014, United States All Rights Reserved.



NETFLIX

Your account is on hold.

Please update your payment details

Hi Dear,

We're having some trouble with your current billing information. We'll try again, but in the meantime you may want to update your payment details.

UPDATE ACCOUNT NOW

Need help? We're here if you need it. Visit the Help Centre or contact us now.

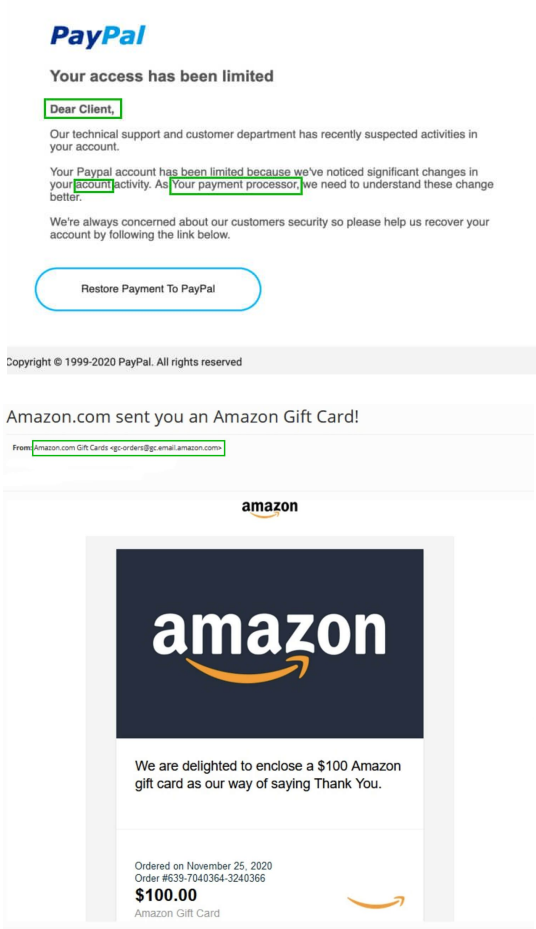
- Your friends at Netflix

Questions? Call 0800 096 8875

This account email has been sent to you as part of your Netflix membership. To change your email preferences at any time, please visit the Communication Settings page for your account.

Please do not reply to this email, as we are unable to respond from this email address. If you need help or would like to contact us, please visit our Help Centre at help.netflix.com.

This message was emailed to [slny]22@sky.com] by Netflix.
ISRC: 12818_10208_1_en-GB_GB
Use of the Netflix service and website is subject to our Terms of Use and Privacy Statement.
©Netflix International B.V.?



PayPal

Your access has been limited

Dear Client,

Our technical support and customer department has recently suspected activities in your account.

Your Paypal account has been limited because we've noticed significant changes in your acount activity. As Your payment processor, we need to understand these change better.

We're always concerned about our customers security so please help us recover your account by following the link below.

Restore Payment To PayPal

Copyright © 1999-2020 PayPal. All rights reserved

Amazon.com sent you an Amazon Gift Card!

From: Amazon.com Gift Cards <gc-orders@gc.email.amazon.com>

amazon

amazon

We are delighted to enclose a $100 Amazon gift card as our way of saying Thank You.

Ordered on November 25, 2020
Order #639-7040364-3240366

$100.00
Amazon Gift Card

Thank You for submitting your claims. This email acknowledges receipt of your details.

The NC COVID-19 VACCINE LOTTERY drawings are part of Gov. Roy Cooper's push to get more people vaccinated against COVID-19.

Currently, 49% of people in Mecklenburg County have had a least one dose of the vaccine.

North Carolina's lottery drawings are performed with a random number generator in which your number was selected.

We have been appointed a personal CLAIMS CONSULTANT to help facilitate your claims ASAP, You are advised to contact your claims consultant immediately with the following recommended info/documentation to initiate your claims.

1. Winning Reference Number
2. Full Name
3. Date Of Birth
3. Address
4. Mobile #:
5. Occupation
6. A Copy of your Identification (Drivers Licence, Passport or State ID)

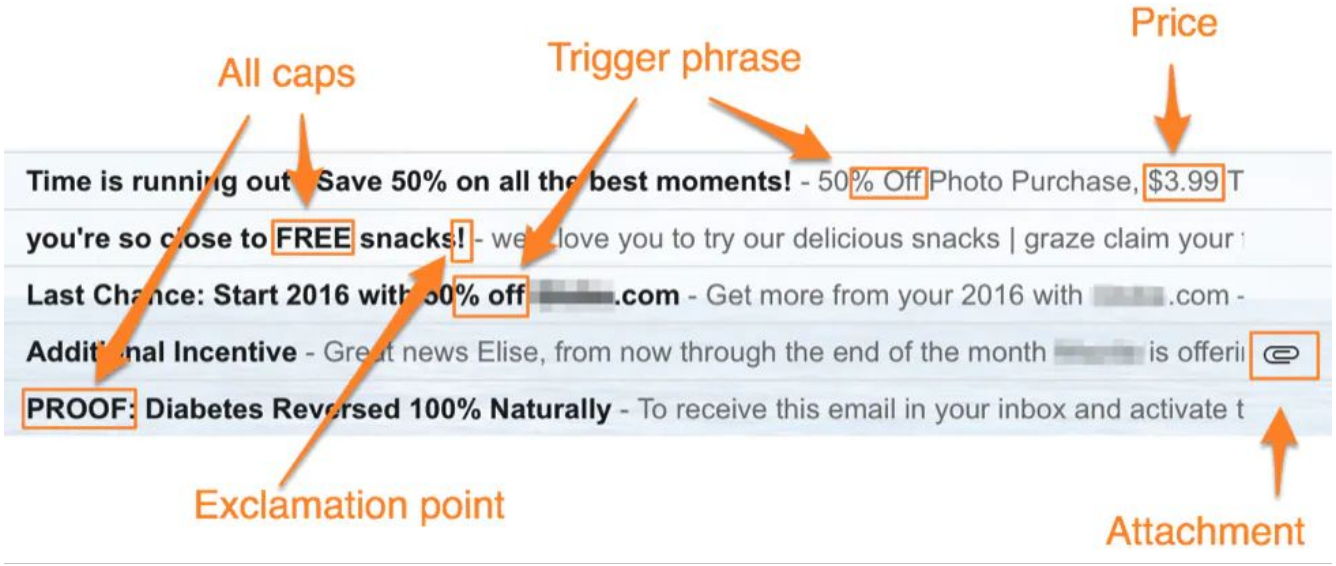
Contact your Clams Consultant below with the above requested informations:

Consultant Name: Ryan Rogers
Consultant Email: ryanrogers56@aol.com

After validation of your details requested above, Your personal CLAIMS CONSULTANT with NCDHHS officials will reach out to you via email or phone. Validatic process can take from 24-72hrs after submitting of your IDENTIFICATION DOCUMENT to your claims concultant.

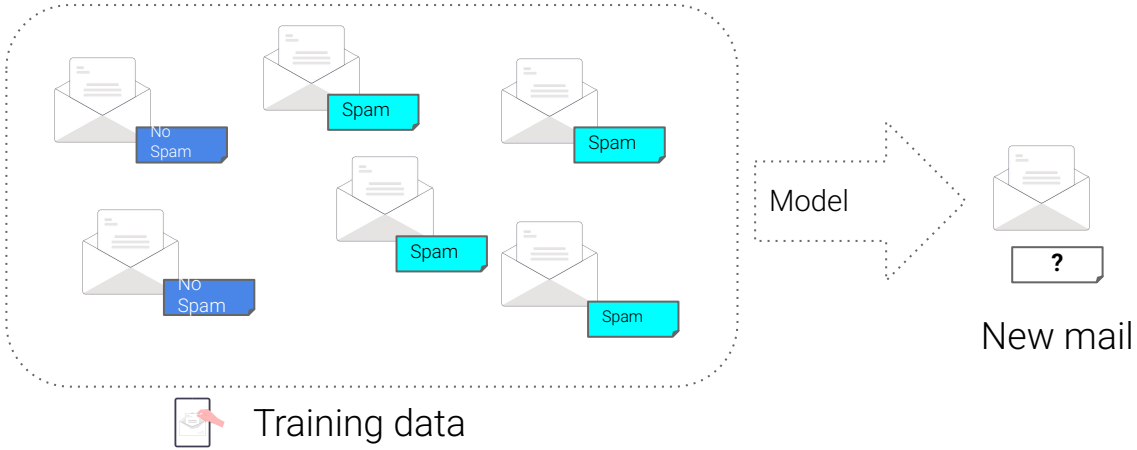
Find more details about NC COVID-19 VACCINE LOTTERY HERE

NC COVID-19 VACCINE LOTTERY.



Picture: Williams (2018)

Predicting spam mails



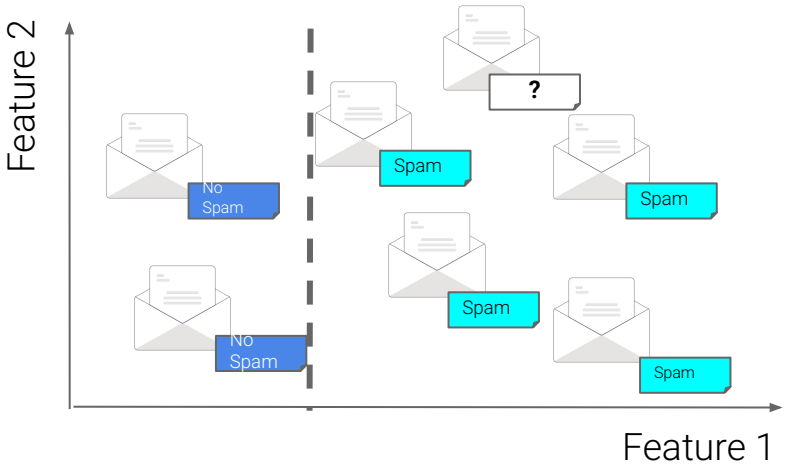
Observation:

Labels: No Spam Spam

Features: # of words, # of capital letters, price included, all caps, # of exclamation points …

Source: Géron (2019)

Predicting spam mails



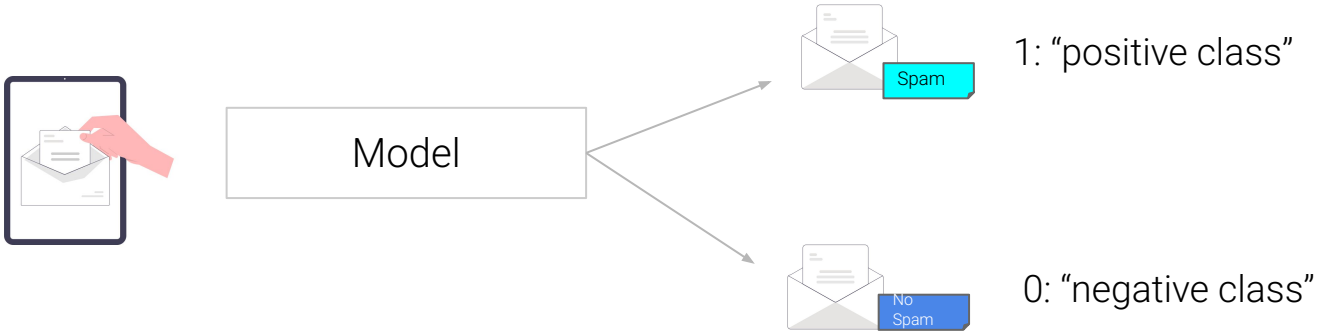
Observation: 

Labels: No Spam Spam

Features: # of words, # of capital letters, price included, all caps, # of exclamation points …

Source: Géron (2019)

Predicting spam mails



Confusion matrix: visualizing the performance of a classification model (binary classification)

	Predicted: 1 Spam	**Predicted: 0** No Spam
Actual: 1 Spam		
Actual: 0 No Spam		

Source: Wilber (2022)

Confusion matrix: visualizing the performance of a classification model (binary classification)

	Predicted: 1 Spam	**Predicted: 0** No Spam
Actual: 1 Spam	True Positive (**TP**)	
Actual: 0 No Spam		

Source: Wilber (2022)

Confusion matrix: visualizing the performance of a classification model (binary classification)

	Predicted: 1 Spam	**Predicted: 0** No Spam
Actual: 1 Spam	True Positive (**TP**)	
Actual: 0 No Spam		True Negative (**TN**)

Source: Wilber (2022)

Confusion matrix: visualizing the performance of a classification model (binary classification)

	Predicted: 1 Spam	**Predicted: 0** No Spam
Actual: 1 Spam	True Positive (**TP**)	
Actual: 0 No Spam	False Positive (**FP**)	True Negative (**TN**)

Source: Wilber (2022)

Confusion matrix: visualizing the performance of a classification model (binary classification)

	Predicted: 1 Spam	**Predicted: 0** No Spam
Actual: 1 Spam	True Positive (**TP**)	False Negative (**FN**)
Actual: 0 No Spam	False Positive (**FP**)	True Negative (**TN**)

Source: Wilber (2022)

Decomposing predictions

True Positives (**TP**): The number of positive instances correctly classified as positive.

True Negatives (**TN**): The number of negative instances correctly classified as negative.

False Positives (**FP**): The number of negative instances incorrectly classified as positive.

False Negatives (**FN**): The number of positive instances incorrectly classified as negative.

Source: Wilber (2022)

Decomposing predictions

True Positives (**TP**): The number of positive instances correctly classified as positive.

- E.g., predicting an email as **spam** when it actually is **spam**.

False Positives (**FP**): The number of negative instances incorrectly classified as positive.

- E.g., predicting an email is **spam** when it actually is **not spam**.

True Negatives (**TN**): The number of negative instances correctly classified as negative.

- E.g., predicting an email is **not spam** when it actually is **not spam**.

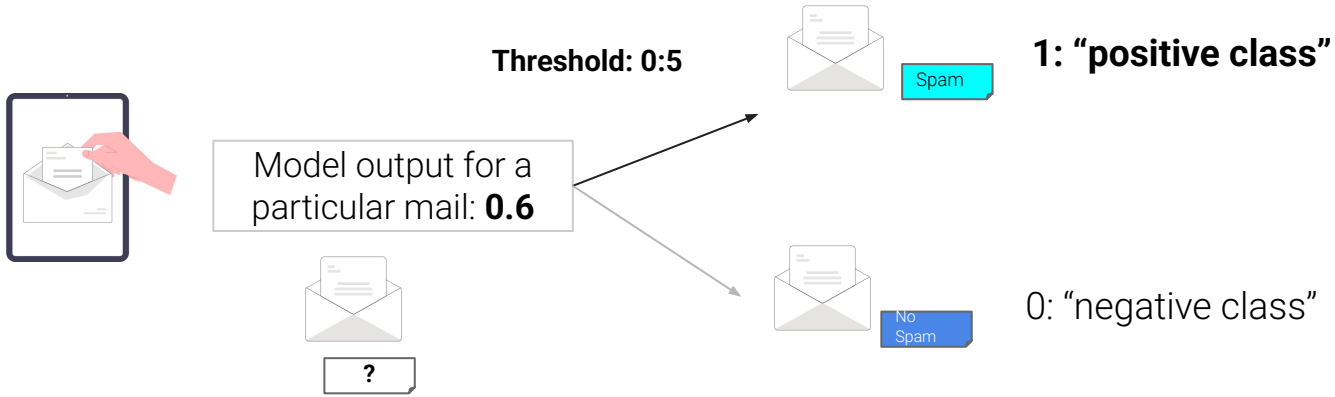
False Negatives (**FN**): The number of positive instances incorrectly classified as negative.

- E.g., predicting an email is **not spam** when it actually is **spam**.

Source: Wilber (2022)

Classification (decision) threshold

- Translate our numeric model predictions (e.g. 0.6) into distinct categories (e.g. **spam** or **no spam**).
- We need to define a **threshold** (e.g. 0.5)

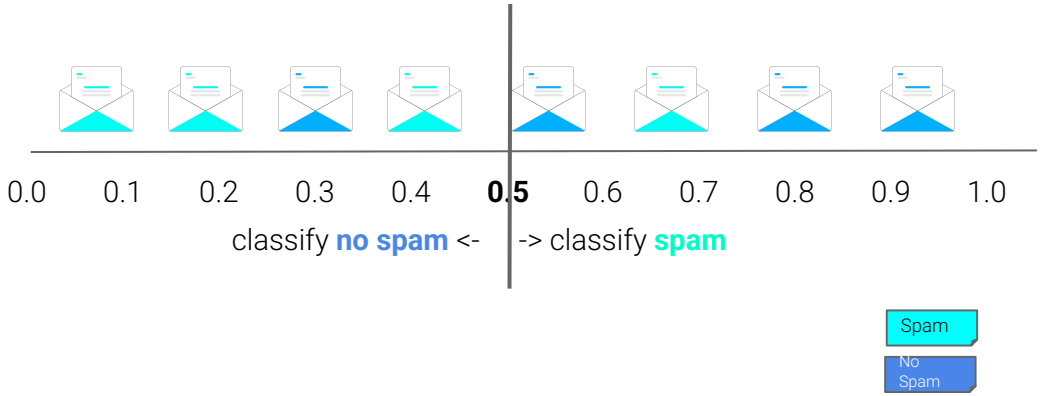


Source: Wilber (2022)

Classification threshold

For example, if our classification threshold is 0.5:

- we'll classify any email with a probability greater than 0.5 as being **spam**
- and any mail with a probability less than 0.5 as being **no spam**



Source: Wilber (2022)

Literature

Çetinkaya-Rundel, M. & Hardin, J. (2021). Introduction to Modern Statistics. Springer. URL: https://openintro-ims.netlify.app/index.html

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly UK Ltd.

Wilber (2022). Precision & Recall - Accuracy Is Not Enough. MLU-Explain. URL: https://mlu-explain.github.io/precision-recall/