

Email Risk using IPQualityScore API

This notebook examines the process of integrating with the IPQualityScore Email Validation API, analyzing the responses, and categorizing emails based on their perceived level of fraud risk.

1. API Integration

The notebook successfully integrates with the IPQualityScore Email Validation API using the `requests` library in Python. A dedicated function `validate_email_ipqs` was created to handle API calls, including constructing the request URL, setting a timeout, and basic error handling for unsuccessful API calls.

Data Format:

- Inputs:
 - API key (create an account at [IPQualityScore Website](#)) and change it to yours
 - Email (to be analyzed)
- Output:
 - JSON with 60 rows ([see full output](#)), the exact definition can be seen [here](#), main fields include:
 - Fraud score
 - Valid
 - Disposable
 - Recent Abuse
 - Spam Tramp Score
 - First seen

2. Fraud Risk Analysis Methodologies

The core analysis is performed by the `process_responses` function, which takes a list of API responses and transforms them into a pandas DataFrame. Key fields from the API response such as `valid`, `disposable`, `honeypot`, `recent_abuse`, `fraud_score`, and `first_seen` are extracted and included in the DataFrame.

The `timestamp_days` is calculated from the `first_seen` timestamp to understand how recently an email address was first observed

A `risk_level` is assigned to each email based on a set of rules:

- **High Risk:** Emails that are not valid, are disposable, are honeypots, or have a fraud score greater than 75.

- **Medium Risk:** Emails with a fraud score greater than 25 or a `timestamp_days` less than 10 (indicating a recently created email).
- **Low Risk:** Emails that do not fall into the High or Medium risk categories.


3. Findings

Based on the sample emails analyzed:

- The majority of the sample emails were categorized as **High Risk**. This is expected given that the sample includes known disposable emails, a known spam email, and an invalid email format.
- A smaller number of emails were categorized as **Low Risk**, which included valid business and personal email addresses.
- One email was categorized as **Medium Risk**, which was the recruiter's email, considering a valid perspective since it should not be used in many applications on the web (without a known fingerprint).

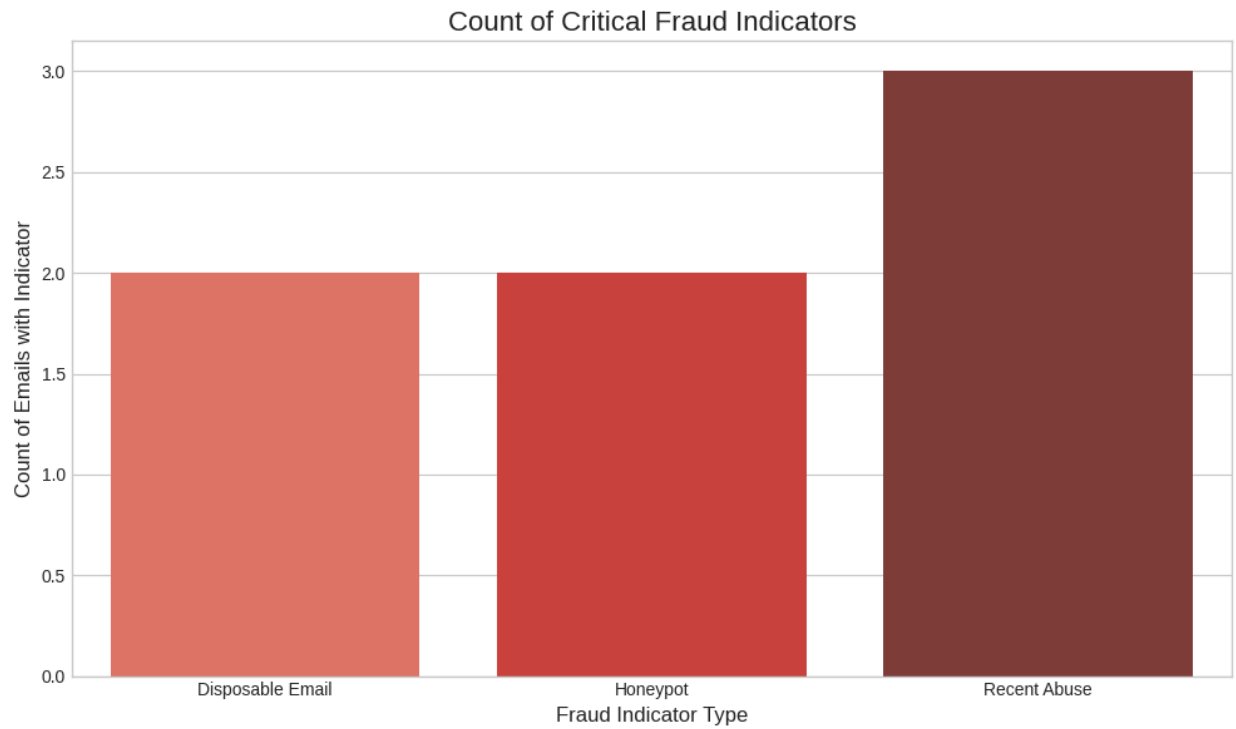
The analysis demonstrates the effectiveness of using the IPQualityScore API in identifying potentially fraudulent or risky email addresses based on various indicators and a calculated fraud score.

Appendix

- Response [parameters definition](#)
-  [Mauro Cesar Risk Analyst API](#)
- [Google Collab Notebook](#)

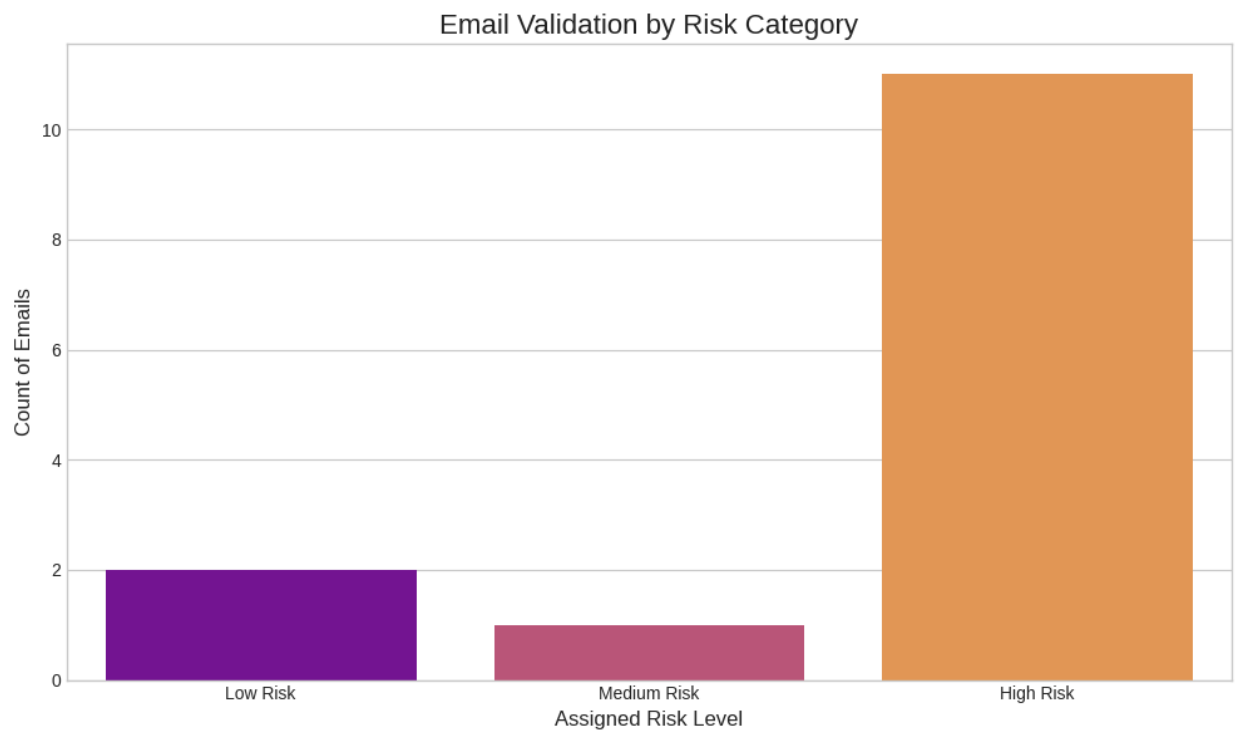
Critical Fraud Indicators

There is a high indication of recent abuse in the sample data (3 customers).



Risk Category Distribution

Most of the sample is categorized as high risk.



Sample output

The output from the code can be seen as below, the segmentations are as expected, which indicate a good risk signal. Increase better evaluate we need backtesting data for a larger sample.

index	email	is_valid	is_disposable	is_honeypot	has_recent_abuse	fraud_score	spam_score	human	timestamp_days	risk_level	Comments
0	test.user@gmail.com	FALSE	FALSE	FALSE	TRUE	100	low	9 years ago	3124	High Risk	Random Generated
1	legit.customer@yahoo.com	FALSE	FALSE	FALSE	FALSE	91	none	19 hours ago	0	High Risk	Random Generated
2	another.one@hotmail.com	FALSE	FALSE	FALSE	FALSE	91	none	9 years ago	3124	High Risk	Random Generated
3	student123@outlook.com	TRUE	FALSE	FALSE	FALSE	80	none	19 hours ago	0	High Risk	Random Generated
4	contact@ipqualityscore.com	TRUE	FALSE	FALSE	FALSE	0	none	2 years ago	550	Low Risk	Business Valid
5	test@trash-mail.com	FALSE	TRUE	FALSE	TRUE	100	low	9 years ago	3124	High Risk	Disposable Example
6	user@10minutemail.com	FALSE	TRUE	FALSE	TRUE	100	none	19 hours ago	0	High Risk	Disposable Example
7	bademail@gmail.com	FALSE	FALSE	FALSE	FALSE	91	none	9 years ago	3124	High Risk	Know abuser GPT
8	thisisnota.valid.email@	FALSE	FALSE	FALSE	FALSE	91	none	just now	0	High Risk	Invalid Format
9	notauser@nonexistentdomain123.com	FALSE	FALSE	FALSE	FALSE	96	none	19 hours ago	0	High Risk	Invalid Format
10	maurocvdm@gmail.com	TRUE	FALSE	FALSE	FALSE	0	none	9 years ago	3124	Low Risk	Mine
11	marilis.redondo@lalamove.com	TRUE	FALSE	FALSE	FALSE	0	none	just now	0	Medium Risk	Recruiter
12	noreply@bydoor.com	TRUE	FALSE	TRUE	FALSE	0	medium	3 years ago	1027	High Risk	Mkt from random app
13	aa4515490@gmail.com	TRUE	FALSE	TRUE	FALSE	95	high	just now	0	High Risk	SPAM I received

Fraud Score Distribution

The distribution is highly concentrated near the boundaries, as expected, since the emails generated are more likely to be fraudulent or personal emails that I know are not.

Distribution of Email Fraud Scores (IPQS)

