Task 1: Improving KYC

1. Overview

Data in doc_reports.csv and facial_similiary_reports.csv span from the near- end of May to the end of October (2017-05-23 15:13:02 to 2017-10-31 23:54:24). Because data for May does not cover the entire month and are therefore few, the month of May will be excluded from the analysis that follows.

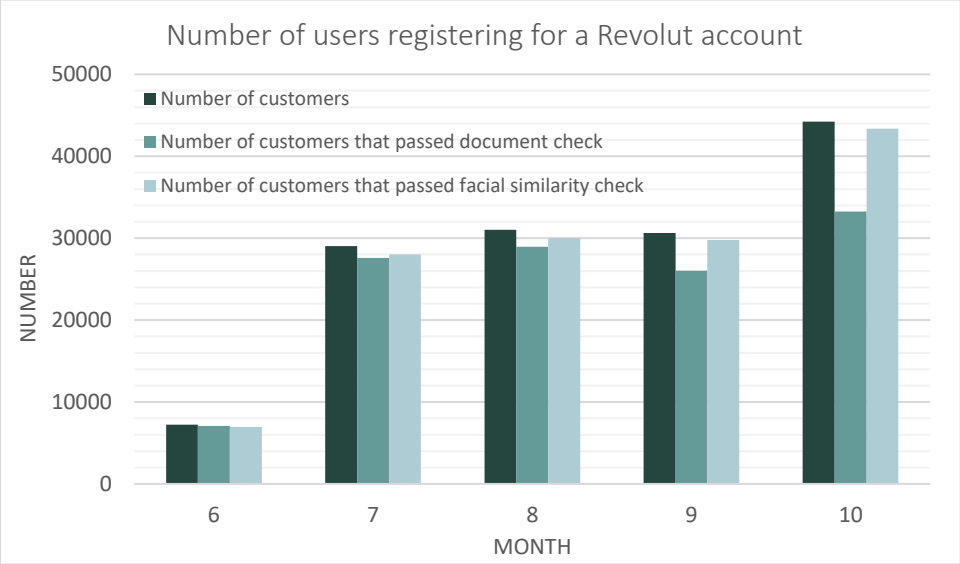

Figure 1: Number of customers who want to open a Revolut account from June to October.

Looking at month-on-month growth, it is found the number of customers that want to open a Revolut account rose rapidly (yay!).

2. Pass rates

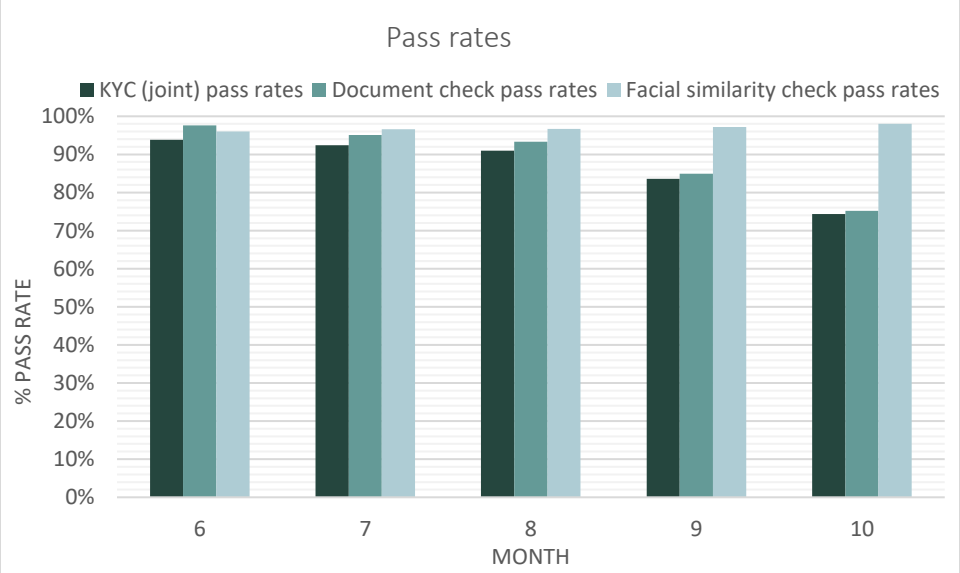

Figure 2: Pass rates for the overall KYC process, for the document check only and for the facial similarity check only. Pass rate for the overall KYC process is also the onboarding rate.

Looking at the overall and separate pass rates, we can see that

- The overall KYC pass rate decreased substantially from 94% in June to 74% in October.
- The document check pass rate fell steadily and significantly from 98% in July to 75% in October.

- The facial similarity check pass rate remained above 95% in all months.

Based on the trends above it is therefore clear that decreasing pass rate in document checks is the primary cause to decreasing overall pass rate.

Now that we have concluded that document check is the root cause to our problem, we can focus on document checks, to see what is leading to a decrease in their pass rates.

3. Document check

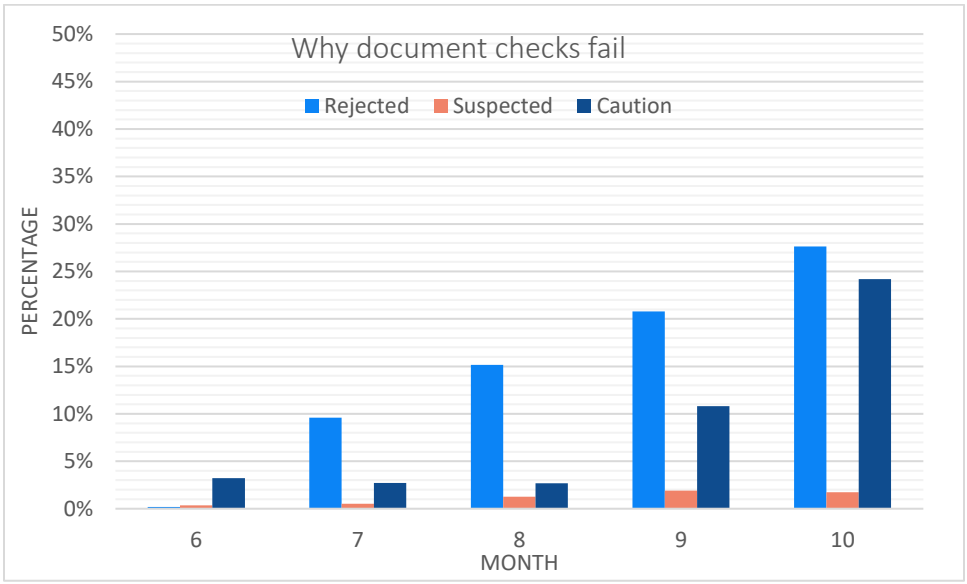

Figure 3: "Rejected", "suspected" and "caution" sub-results. The percentage value is calculated out of number of customers who attempted the process.

The results of document checks can be categorized as "clear", "rejected", "suspected", or "caution". It can be seen from Figure 3 that out of all document checks:

- "Rejected" is the main reason for document check failures.
- The percentage of "rejected" cases tripled from July to October, from 10% to 28%.
- While the percentage of "caution" cases remained steady in June to August, it exploded in the last two months. "Caution" cases became as significant as "rejected" cases in October.
- The percentage of "suspected" cases is small, at 1% to 2% in all months.

3.1 "Rejected" document checks

First, we shall focus on the "rejected" cases. Document checks are flagged as "rejected" when the report has returned information where the check cannot be processed further. Because the number of "image_quality_result" and "supported_document_result" being "unidentified" corresponds to the number of "rejected", we can confidently deduce that these sub-breakdowns are the sole criteria that contribute to "rejected" checks.

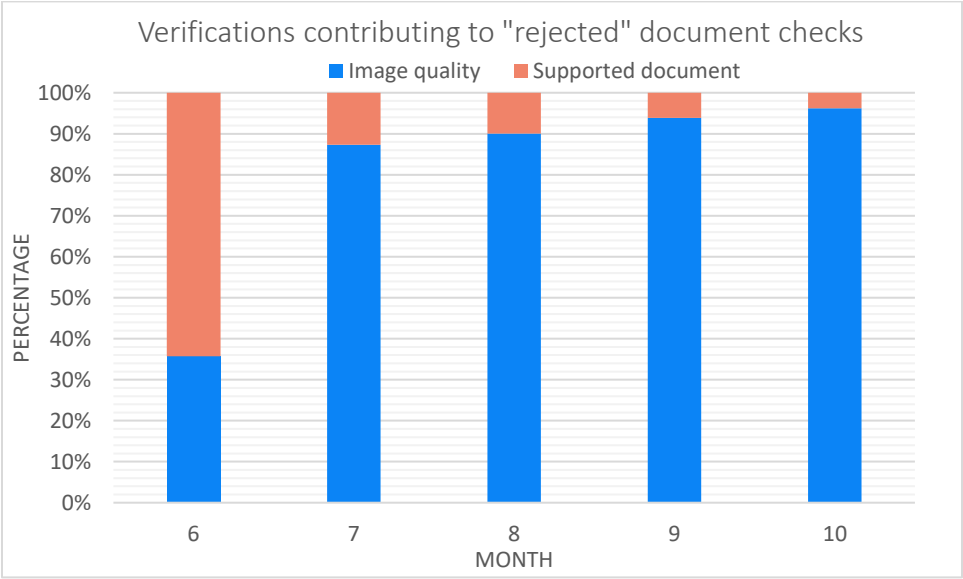

Figure 4: Percentage of 'rejected' attempts with result 'unidentified' in the image_quality and supported_document sub-breakdowns. Percentage values are calculated over number of 'rejected' attempts.

Looking at the contributions of image quality and supported document reports, it is found that around 90% of "rejected" document checks is due to poor image quality. It can therefore be concluded that image quality is the primary verification that causes document checks to be "rejected".

3.2 "Caution" document checks

Now that we have tackled the "rejected" document checks, we shall move on to the second cause of document check failures --- "caution". Document checks are flagged as "caution" if document can be processed, but image quality is too low to make a fraud assessment.

It turns out (from summing the number of cases) that a number of verifications, namely "visual_authenticity", "image_integrity", "data_validation" and "data_comparison has led to documents being classified as "caution".

Month	6	7	8	9	10
Visual_authenticity	118	484	430	235	156
Image_integrity	0	0	196	2943	10299
Data_validation	59	290	239	220	393
Data_comparison	59	0	0	0	0

Table 1: Number of 'caution' attempts with result 'consider' in the visual_authenticity, image_integrity data_validation, and data_comparison reports.

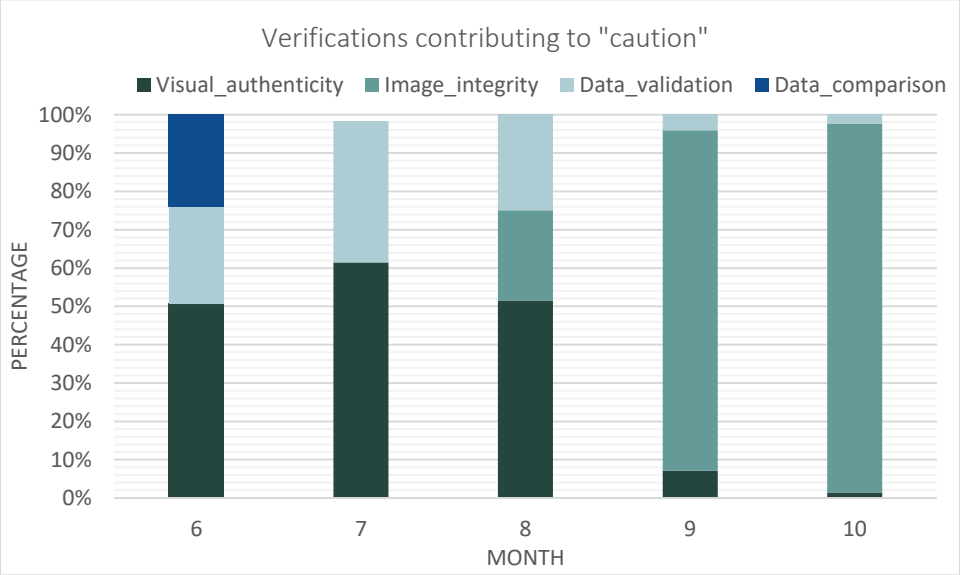

Figure 5: Percentage of 'caution' attempts with result 'consider' in the visual_authenticity, image_integrity data_validation, and data_comparison breakdowns. Percentage values are calculated over number of 'caution' attempts.

Table 1 shows that:

- The number of failures in visual authenticity fell steadily from July to October.
- The number of failures in data validation remained virtually constant from July to October.
- The number of failures in data comparison dropped from 59 in June to 0 in July, and remained at 0 since.
- In contrast, the number of image integrity failures ballooned from 0 in July to 10299 in October.

Analysing from a percentage perspective, it can be seen from Figure 5 that

- Image integrity became the dominant contributor to "caution" cases in the last two months.

Recall that previously we saw an alarming climb in "caution" results in September and October. Based on the data above, it is now clear that this development is caused by a sudden increase in failed image integrity reports.

Once we have established that image_integrity is the primary cause of "caution", we shall look at the reasons behind failed image_integrity reports.

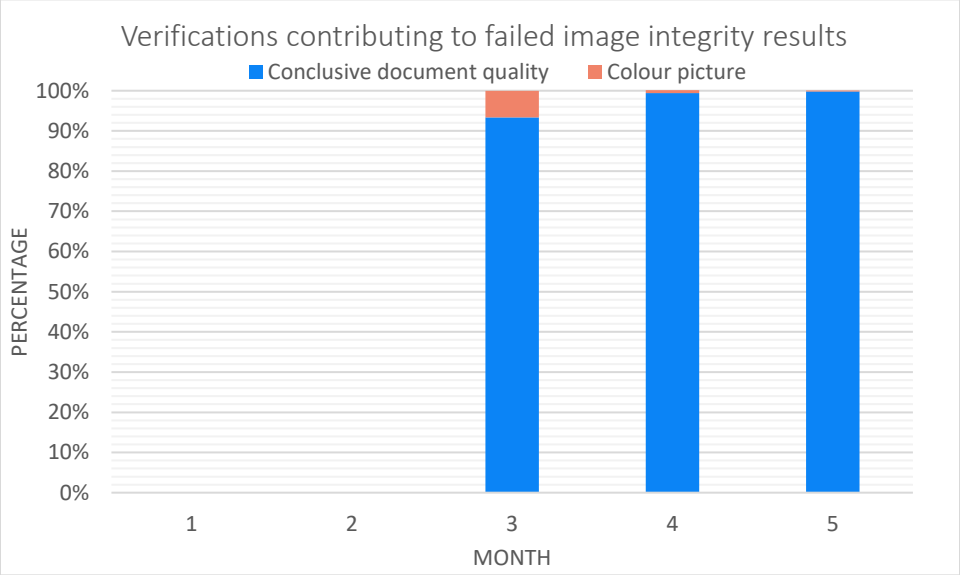

Figure 6: Percentage of 'caution' attempts with result 'consider' in the conclusive_document_quality and colour_picture sub-breakdowns. Percentage values are calculated over number of unidentified image_integrity attempts.

As seen in Figure 6, more than 90% of failed image integrity results is due to failed conclusive document quality . In conclusion, the dramatic increase in document check results with “caution” in September and October is due to an enormous increase in failed conclusive document quality.

4. Solution

In Section 2 and Section 3, it is found that failed document check is the root cause to decreasing KYC pass rates in the recent period. The main reasons to failed document checks are:

- Rejected: Failed image quality
- Caution: Failed conclusive document quality

‘Image quality’ is insufficient when:

- A low-resolution image where the document information is not readable.
- The MRZ (Machine Readable Zone) is obscured/unreadable.
- Vital data points are obscured/unreadable.

‘Conclusive document quality’ can fail due to the following reasons:

- Obscured data points, e.g. date of birth obscured
- Obscured security features, e.g. glare obscuring a security chip
- Abnormal document features, e.g. severely diminished colours
- Watermarks or text overlay, e.g. watermark on document

These verifications essentially point to a similar problem: the quality of the document uploaded by the user, when they try to submit their government-issued photo ID. Because users only register for a Revolut account through the mobile app, it is highly likely that they will take a picture of their ID, rather than uploading a scan file. This can lead to many of the problems mentioned above.

There are many ways in how an image can be of poor quality, depending on how the user takes the image of the photo ID: they can either be blurred, low resolution, obscured, cropped, held at an unreadable angle, or even be the wrong format that Veritas takes (jpg, png and pdf).

One quick and easy solution is to provide clear instructions on the app on the image requirements. Give guidelines on the image size, resolution and example photos that will not be accepted.

Customer drop-out can also be reduced with the glare detection feature. Strong light shining on an ID causes glare which may obscure essential information. The glare detection feature in the iOS SDK notifies users that their photo contains glare as soon as they take it.

The other more costly, but more secure solution is to develop and run machine learning technologies on the app to detect poor image quality. Images that with poor quality, invalid format etc. should be rejected immediately and users should be prompted to take another photo.

Please see Appendix A for data inconsistencies.

Appendix A: Inconsistencies

Several problems can be found in the data:

1. There are a lot of duplicate entries.
   - In each of doc_reports.csv and facial_similarity_reports.csv, the number of duplicate entries is 4252 (Appendix B1,a).
   - Over which the number of duplicate entries that failed is 1123 (Appendix B1,b).

2. Although it is specified that each customer has up to 2 attempts, it is apparently not the case.
   - The number of people who attempted more than 2 times, at different times is 264 (Appendix B2,a).
   - Over which the number of customers that attempted more than 2 times even though they failed both times is 45 (Appendix B2,b).

3. There is a bizarre attempt_id at row number 174083, user_id 7563e.
   - The attempt_id is recorded as 9.69E+31 in doc_reports.xls (Appendix B3).
   - The attempt_id is recorded as 9.6949E+31 in facial_similarity_reports.xls
   - User ids are usually recorded as a string that consists of 32 hexadecimal digits. However, in this unique case, because all 32 characters in the string are numbers, the Excel software automatically converts the data into scientific numerical format.
   - As a result of this problem, the attempts in doc_reports.xls and the attempts in facial_similarity_reports.xls are not 1 to 1.

4. In the month of July, 20 non-duplicate document check attempts failed with 'caution', even though all underlying verifications are either `clear` or blank (Appendix B4).

Appendix B: SQL queries

Queries used in this task follow a similar logic and therefore have similar styles. For example, to count the number of customers with document check 'clear' in each month:

```
SELECT COUNT(DISTINCT user_id)
FROM doc_reports
WHERE result = 'clear'
GROUP BY strftime('%m', created_at);
```

1.

(a) Number of duplicate entries in doc_reports

```
SELECT COUNT(*)
FROM (SELECT dr.user_id, dr.created_at, dr.result, COUNT(*)
    FROM doc_reports dr
    GROUP BY dr.user_id, dr.created_at, dr.result
    HAVING COUNT(*) > 1);
```

(b) Number of duplicate entries in doc_reports that failed

```
SELECT COUNT(*)
FROM (SELECT dr.user_id, dr.created_at, dr.result, COUNT(*)
    FROM doc_reports dr
    WHERE result = 'consider'
```

```
    GROUP BY dr.user_id, dr.created_at, dr.result
    HAVING COUNT(*) > 1);
```

2.

(a) Number of customers who attempted more than 2 times

```
SELECT COUNT(*)
FROM (SELECT dr.user_id, dr.created_at, dr.result, COUNT(*)
    FROM doc_reports dr
    GROUP BY dr.user_id, dr.created_at, dr.result
    HAVING COUNT(*) > 2);
```

(b) Number of customers who failed more than 2 times

```
SELECT COUNT(*)
FROM (
    SELECT user_id, created_at, count(*)
    FROM (SELECT DISTINCT dr.user_id, dr.created_at, dr.result
            FROM doc_reports dr)
    WHERE result = 'consider'
    GROUP BY user_id
    HAVING COUNT(*) > 2
);
```

3. Attempt_id that is not mutual in doc_reports.xls and facial_similarity_reports.xls

```
SELECT number, dr.attempt_id
FROM doc_reports dr
EXCEPT
SELECT number, fsr.attempt_id
FROM facial_similarity_reports fsr;
```

5. Distinct attempts that has sub_result caution, even though tests are either `clear` or are blank.

```
SELECT DISTINCT user_id, created_at, sub_result
FROM doc_reports
WHERE sub_result = 'caution'
    AND (visual_authenticity_result != 'consider'
    AND data_validation_result!= 'consider'
    AND data_comparison_result !='consider'
    AND image_integrity_result != 'consider'
    AND data_consistency_result !='consider'
    AND police_record_result != 'consider'
    AND compromised_document_result != 'consider');
```