Task-1: Improving KYC

1. Comments on Data Quality:

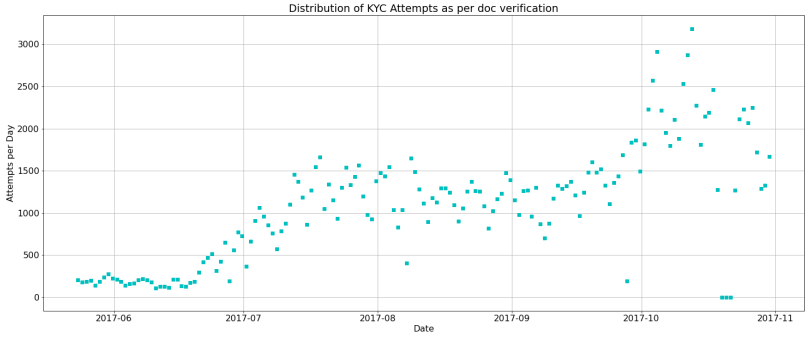
Before we dive into the analysis, let us first talk about the quality of the data that was provided. Unfortunately, the data presented was most likely corrupted because it didn't follow many constraints that were set up as a part of the KYC process. To give a couple of examples:

1. For every user_id, there are two attempts. However, there were exactly 1232 user_ids which had more than 2 attempts. In addition to that, there were also users who had undertaken a second or further attempt in spite of passing the verification process in the first try.
2. As per the Veritas API, gender can only take values male and female (in lowercase). However, not only were there higher case versions of them, but also, a numerical gender value of "8" was found in the data.
3. Country ISO codes are meant to be three "letter" codes as per the API. However, many countries were coded in their digit ISO equivalents. This meant some countries had digit as well as letter representation. Our report doesn't convert the digit ISO into letter ISO, so some reports of a country may be attributed to a separate group which is actually the same country to begin with.

These suggest that the data might either be corrupted (as mentioned earlier) or doctored. Or on a more reasonable level, it is also possible that Veritas' KYC verification system is not robust enough to strictly allow only accepted values in the appropriate fields. Assuming that this is the only data we have, let us move forward.

2. Setting of the problem

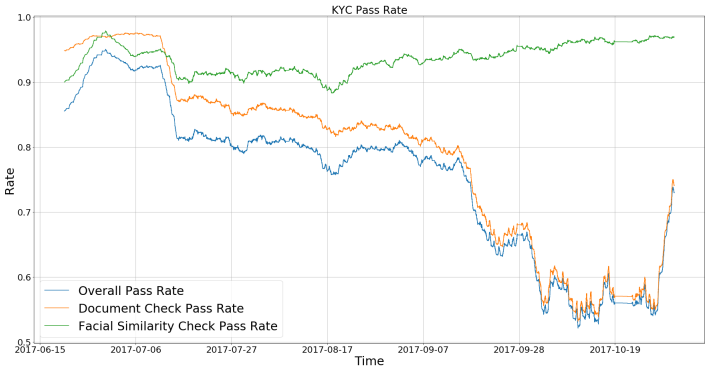
Recently the pass rate for the KYC process has gone down substantially. At the same time, via data analysis, we can see the total number of attempts has increased.



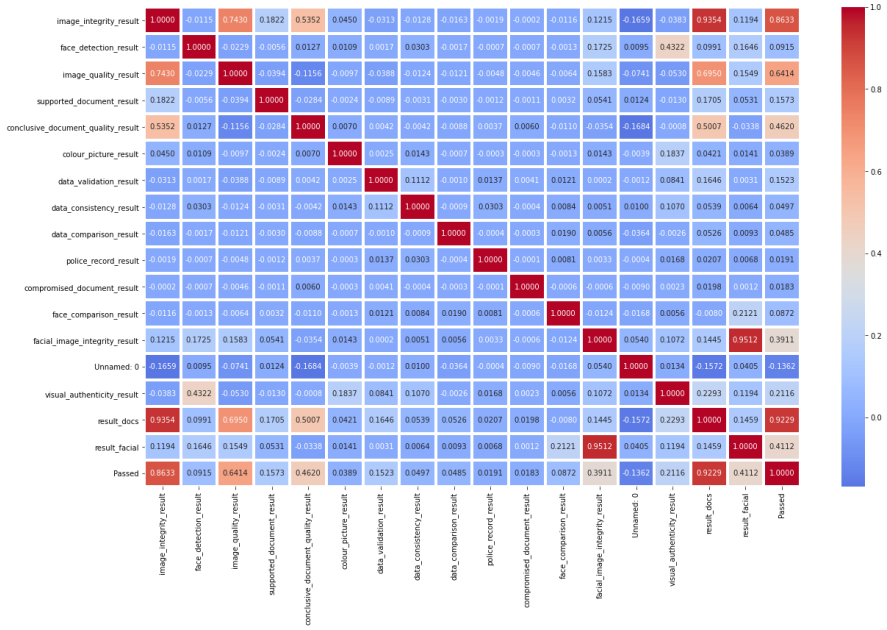
We need to find the possible reasons for the same and suggest solutions that may resolve the problem.

3. Preliminary Analysis of the Data

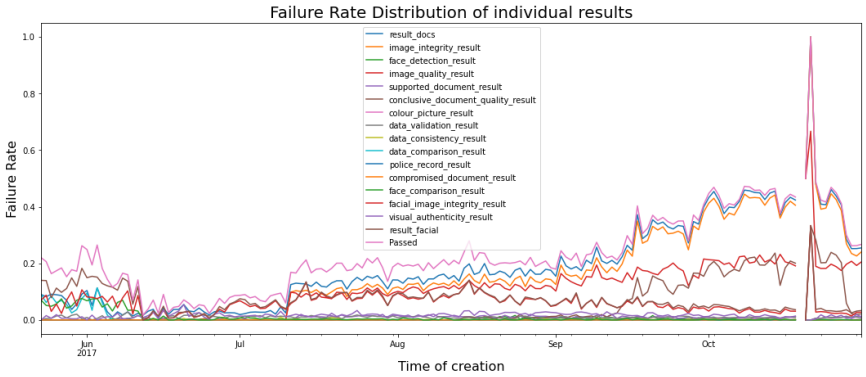
As per the graph shown below, we can see that the overall pass rate has declined in the same fashion as the document check rate has declined. The pass rate for the Facial Similarity Reports seem to have remained stable over the course of the analysis.



The document check report itself consists of sub reports like the image_integrity_result (IIR), the conclusive_document_quality_result(CDQR) and image_quality_result(IQR) among others. We created a heatmap showing the correlation between document check failures and other sub reports.



As can be seen, IIR, CDQR and IQR were leading factors correlated with the document check result. We also plotted a time series graph of all the sub reports and the main reports to observe trends as given below:



From this plot, we also identified some other reports, which while were not highly correlated with the docs_result or the facial_result, had spikes post September. Hence, in addition to the three sub reports previously mentioned, we also analysed facial_image_integrity_result(FIIR).(The subplots can be found in the addition analysis report).{Charts A.1 to A.4}

For all these four sub reports, we dived deeper and tried to find if there were any correlations with the document properties, because as per the API, these reports are indicators of the submitted documents. So, we wanted to see if any specific issuing country/document type/nationality/gender was giving a higher failure rate than the others.

To do the analysis of the sub reports, we needed to find the problem period. To find the problem period, we used the **ADFuller Test** from stats_models to first find the stationarity of the time series for each month and we also simultaneously found the average failure rate over the month. An average failure rate of more than 20 percent within a stationary series could be classified as problematic. Also, a non stationary series with an

average around 20 but an upward trend can also be classified as problematic. This period happened to be the months of September and October of 2017. Even from the graphs above, we can see a rising trend starting from September.
The analysis for each sub report was done for two phases, one in the problem phase (i.e. from 01-09-2017 to 31-10-2017) and one in the non problem phase.

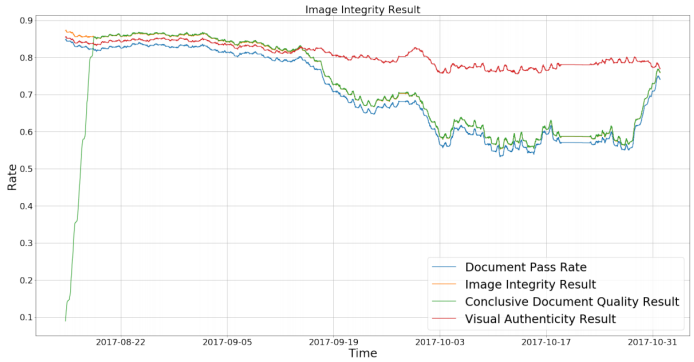
4. Deeper Dive into the spiked sub-reports

The complete analysis (including all graphs) of each sub report for the relevant variables of gender, document_type, issuing_country and nationality is available in the additional analysis report. (Subsection B of the calculation and additional analysis document)

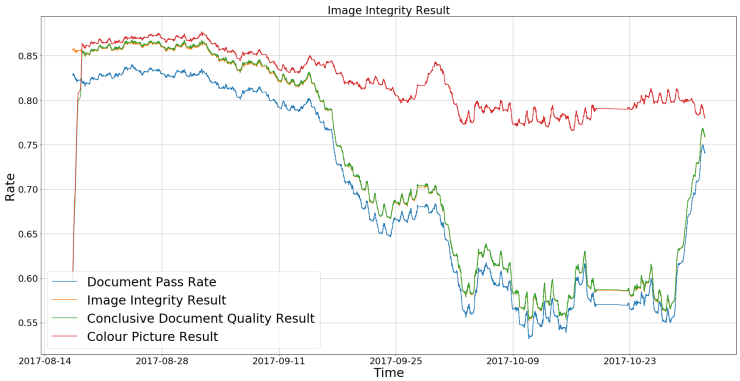
Based on the analysis, we could not find any particular variable amongst document_type, nationality, gender and issuing_country that had a greater influence in any of the sub reports. Based on the magnitude of change and also the trend observed in relation to the document check and facial check result, it was concluded that IIR and CDQR were the most influential in the document check which in turn was the most influential when it came to the pass rate for the KYC. Hence the sub reports of interest going further would be IIR and CDQR.

5. Further Observations with regards to the API, the dataset and financial regulations

The API specifies that if the various fields of address are fit into one field like the “street”, it may lead to lower match levels in the document checks.Also, as per the API, the IIR is made up of the Supported Document result, Color Picture Result, IQR and the CDQR. For the CQDR, the API says A result of `clear` in the `conclusive_document_quality` breakdown of `image_integrity` will assert if the document was of enough quality to be able to perform a fraud inspection. Also, as per the changelogs, there were two major changes, the first one on 14th July’17 to add a new token authentication scheme for the web SDK; and the second one being on 25th Aug’17. The second one was on facial similarity reports and was inconsequential. However, the same can’t be said for the first one.

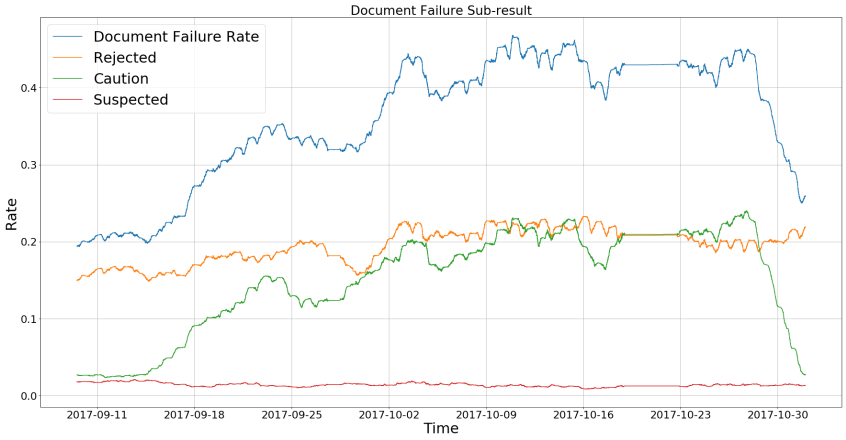


As per the above graph, visual authenticity doesn’t usually fail while the CDQR directly influences the IIR.



From the graphs above and the dataset provided, it can be concluded that colour picture result and CDQR were started to be captured around the second week of August. Although the color picture result was not

influential as much, the CDQR heavily influences the IIR and seems to be very highly correlated with the IIR which in turn influences the pass rate via the document check result.Apart from these sub reports, no other result was influential in the document check result as shown in the graph in the additional analysis report. Finally, let us also see the trend of the sub_result for the document check result.



As per this graph, a significant percentage of the failure rate for the document check during the problematic period can be attributed to the caution sub_result which indicates that the document may not be necessarily fraudulent. **As per news reports, one of the major financial regulations implemented was the money laundering regulation in June 2017 in the UK and some EU countries.**
Finally, no trends were found relating to the nature of the attempts, such as a rise in applicants from a specific nation coinciding with document failure. Neither was there an increase in fraudulent attempts. Therefore, it seems safe to assume that the nature of the attempts did not change.

6. Some Interpretations

The problem seems to stem from the Veritas API, which has clearly been undergoing changes: no properties for the Facial Similarity Checks were recorded in the later months; the Compromised Document Result only started recording results in the latter period; and, color picture and CDQR only started recording results in mid August. These could have been new metrics with the CDQR being changed early on leading to IIR failures and by virtue, document check failures. Furthermore, there could have been other changes which were apparently not significant enough as per Veritas to include into the changelogs but were nevertheless impactful in changing the KYC pass rate. It is therefore necessary to get a more detailed changelog to understand other possible changes that might have concretely influenced the rates.

7. Some probable causes

a. Because MLR 2017 was implemented in June of 2017, it is possible that Veritas made internal changes in an attempt to make their system more robust. However, this might have led to the increase in false positives which has made a dent on the potential user base of Revolut, reducing the profits the company makes from the transactions.
b. There might be logical errors in the code written during an update which may have the possibility of affecting the entire system in ways which are not easily traceable and may start showing effects only after certain inputs have been processed or a certain part of the application process is completed.
c. Since the number of users increased over the months, it is possible that the server on which the Veritas application operates or the Veritas API itself which coordinates the process and provides the interface to the user may be behaving differently because of which the rate fell.

8. Additional Assumptions made for this report

In addition to the initial assumption stated in the “Comments on the Data Quality” subsection, here are some other assumptions that are pivotal to the data analysis process:

a. There has been no error in the translation of information from the servers to the actual CSV files that were provided for analysis in the end.
b. Since we do not have access to previous data, it is assumed that the initial starting points and trends the data shows for the initial few weeks is a stable residue of prior processes that have been verified to not affect the KYC pass rate in an unwanted way.
c. As is a necessary condition for all analytics exercises, it is assumed that the data that is provided is representative of the entire set of reports. In terms of statistical analysis, the sample is representative of the population and is unbiased.

9. Some additional secondary considerations

While the probability of the following points is low, they are worth considering in case the solutions (suggested in the next subsection) don't generate any results.

a. Veritas has different customers operating in different business sectors and different countries. It may be possible that some regulation in a sector different from the BFS may have prompted Veritas to make changes in their overall application. Even regulations from countries where Revolut doesn't operate may have also affected the process in case Veritas doesn't tailor make the application process for each customer.
b. Since MLR 2017, it is unlikely but not impossible that actual fraudsters may be desperately trying to open a Revolut account and siphon off their funds as soon as possible. This may lead to higher fail rates albeit because of true positives instead of false positives.(Positive in the sense it is a fraudster)
c. The internal Computer Vision Algorithm may be a self learning algorithm in the background which may have changed in unexpected ways because of any of the previous reasons or a different unknown reason because of which results have been altered to a great degree and the KYC pass rate reduced.
d. In case Veritas operates the application on different servers, it is possible that some of the servers in a particular location may not be working properly because of a host of reasons and consequently, the received data may be corrupted and fails the test.
e. This is totally unlikely but nevertheless, we can consider the possibility of doctored data entering the system. Since Revolut is considerably ahead of all its competitors, it is possible that a specific competitor or a group of them may be aware of the Veritas API to its core and is thus exploiting its gaps to continuously send synthetic user data that is skewing the kyc rate and making it look as if there is an issue with the process so that Revolut wastes its time and resources on a non existing problem.

10. Suggested solutions

a. Work closely with the Veritas technical team to understand the changes in the API, the non availability of CDQR initially and then the reduction in it's pass rate in the problematic months.
b. Improving the operation time of the process so that applicants with caution or consider statuses get more inclined to resubmit their documents. Also, a more detailed feedback on what the exact problem is would help them make a better submission the next time.
c. Since there is some evidence that there might be inefficiencies from Veritas' end, it would be prudent to search other KYC solution providers who may be more ready to make a tailor made accurate solution for Revolut.
d. For the "caution" status of CDQR, understand what might be the changes that may have prompted the higher rate. Communicate clearly to the Veritas team that the "caution" status leads to loss of business and may lead to the termination of their contract.
e. Understand if any changes have been made to the computer vision algorithms and if it is possible to revert them back.
f. Obtain information on whether other customers of Veritas are facing similar issues or not. In case they are, insist on changing the KYC process.
g. Delegate a similar analysis task to the Veritas team to inspect other features of their entire process from their end and come up with possible solutions.