Data Quality Report – Initial Findings

1. Overview

This report will outline the initial findings based on the cleaned dataset (CreditRisk-7578598_1-2_cleaned.csv). It will summarise the data, describe the various data quality issues observed and how they will be addressed. Please see appendix for some background to this dataset. Appendix includes terminology, assumptions, explanations and summary of changes made to the original dataset. This also includes feature summaries, histograms and boxplots used to visualise the data.

On first indication the dataset appears relatively clean. There are no null values, duplicate columns or columns with irregular cardinalities. There were a small number of rows with no data that have been removed (count 56).The main issues observed were regarding special values for continuous data and numerical scales used for categorical data. In addition, a significant number of outliers were present. Also, several logical tests were carried out on the data a significant number of inconsistencies were found.

2. Summary

Several tests were carried out to check the logical integrity of the data. This brought about a significant number of failures of the data. In total 184 instances of irrational data was observed. For example, in 74 instances the number of satisfactory trades was found to be greater than the total number of trades. This is clearly impossible. This irrational data will need to be dealt with and should be checked with the domain expert. See logical integrity section for further details.

For the continuous features there was the inclusion of several special values i.e. negative values which map to a specific meaning. -9 means “no data”, -8 means “no valid data” and -7 means “condition not met” (see appendix). These values need to be addressed as they are appear in features that cannot logically contain negative numbers. Special value -8 should be changed to a “null value” so that the rest of the useful data can be kept. These values should then be evaluated on a feature by feature basis to see if imputation is possible, i.e. mapping to the median value. The -7 value is a positive marker representing the best possible case and therefore should be set as the largest value in the features that they appear.

For the categorical values several changes are recommended. There are 2 main features both measuring if an entry has been delinquent or not. One feature looks at the last 12 months and the over the total lifetime. They both use different scales that should be mapped to 1 common scale to allow easier comparison and interpretation. In addition, the meaning of “unknown delinquency” should be changed as it likely represents a “null value”. This should be checked with the domain expert.

There was a significant number of outliers present across the feature set. However, on first indication these values appear to be plausible but should be investigated further.

3. Review Logical Integrity

13 tests were carried out. The failures are below;

- Test 1 - Check if any entries have number of satisfactory trades > than number of total trades
  - 74 cases found
- Test 3 - Check if any entries have number trades open in last 12 months > number of total trades (impossible)
  - 11 cases found
- Test 5 - Check if any entries have number trades 90 days late > number of total trades (impossible)
  - 4 cases found
- Test 6 - Check number months since most recent trade > number months since oldest trade (impossible)
  - 31 cases found
  - Most failures due to -8 special values which will be handles separately
- Test 7 - Check if any entries have number revolving trades with balance > number total trades (impossible)
  - 13 cases found
- Test 8 - Check if any entries have number install trades with balance > number total trades (impossible)
  - 11 cases found
- Test 9 - Check if any entries have number bank trades with high utilization > number total trades (impossible)
  - 2 cases found
- Test 10 Check for entries that have no trades but entry for months since trade open (impossible)
  - 9 cases found
- Test 11 Check for % trades never delinquent == 100% and have a trade over 60 days late (impossible)
  - 26 cases found
- Test 12 Check for % trades never delinquent < 100% and have no trade over 60 days late (impossible)
  - 234 cases found
  - Check with domain expert - we do not have any feature for number of trades 30 days late which could account for this. No action will be taken as there are too many rows effected
- Test 13 Check for months since most recent delinquency == 0 and have no trade over 60 days late (impossible)
  - 3 cases found

4. Review Continuous Features

4.1. Descriptive Statistics

There are 22 continuous features. All continuous features can be grouped into 1 of 7 main categories which will be summarised below;

- External (Count 1)
    - Risk Estimate (Scale: 0-100) - Without knowing how it is calculated the values above seem plausible. The majority of value fall in the range (65-80). There are a few outliers in the 42-52 range, which would be the highest risk. Overall the distribution appears normal.
- Months Since (Count 4)
    - The special value "-8" appears in 3 out of 4 of these features, indicating no data available and will need to be addressed.
    - "Months Since Most Recent Delinquency" has over 40% of values with the special value "-7" meaning condition not met (i.e. never delinquent) and will need to be addressed.
- Average Months in File (Count 1)
    - "Average Months in File" has several outliers due to old accounts. All seem plausible but should be investigated further.
- Number of Trades (Count 8)
    - Al features measure the number of trades against a specific criteria. 3 of these features (Number Revolving Trades with Balance, Number Install Trades with Balance, Number Bank Trades with High Utilization) contain the special value "-8" and will need to be addressed.
- Number of Inquiries Made (Count 2)
    - Both features measure number of inquiries made in last 6 months, however one excludes the last 7 days. Both features have very similar values as expected and make sense in relation to each other.
- Percentage of Trades (Count 3)
    - All features measure the percentage of trades against a specific criteria.
    - The feature "percentage trades with balance" has 1 entry with special value -8 and will need to be addressed.
- Net Fraction of Trades (Count 2)
    - All features measure the net fraction of trades against a specific criteria. Net fraction is measured out of 100 and so is analogous to percentage. Both features contain the special value -8.
    - NetFractionInstallBurden has over 300 entries with -8 and represents over 30% of all entries, while NetFractionRevolvingBurden has 13 entries. Both will need to be addressed.
    - It is interesting to note that both features have a small number of entries with net fractions over 100, indicating the entries owe more credit they have been allocated. This value warranty a check with the domain expert to ensure this interpretation is correct.

Many of these features have outliers. All seem plausible but should be investigated further.

4.2. Histograms

All histograms can be found on the appendix as summary sheet. Individual plots can be found in the accompanying notebook. As there are over 20 features an in-depth review will not be carried out here. Overall the features showed plausible distributions. The outliers will be investigated further but no immediate action expected.

4.3. Box plots

All boxplots can be found on the appendix as summary sheet. Individual plots can be found in the accompanying notebook. Again, outliers will be investigated further but no immediate action expected.

5. Review Categorical Features

5.1. Descriptive Statistics

There are 3 categorical features in the dataset, 1 of which is the target and will not be evaluated here. The 2 remaining are “Max delinquency ever” and “Max delinquency over the last 12 months”.

Both features use a numerical value which maps to a specific meaning. A lower number indicated a worse meaning. It is interesting to note that the scales used are slightly different, however could be made equivalent. It is recommended to convert the “Max delinquency over the last 12 months” scale to match the “Max delinquency ever” scale as it is easier to interpret. The reason for this is the “Max delinquency ever” scale has single number for each meaning, whereas “Max delinquency over the last 12 months” uses 2 numbers for a single meaning in some cases.

Aside from the different scales, we can see that for “Max delinquency over the last 12 months” there is a significant amount of values in the “unknown delinquency” range (index 7), with approximately 30% of all values. This is not expected. Since these values are from the last 12 months i.e. more recent, it should not be difficult to find out if there was a delinquency or not during that time. I suspect that “unknown delinquency” is the equivalent of a “null value”. This should be checked with the domain expert. It is likely that a delinquency event would not be omitted. Therefore, in absence of any further information, I would recommend a “null value” is treated the same as “current or never delinquent” and should be mapped as such.

For “Max delinquency ever” there are far fewer “unknown delinquency” values (6 in total), which seems more reasonable. Again, in the absence of input from the domain expert I would recommend this value to be mapped to “current or never delinquent” to keep both features aligned.

The category “unknown delinquency” will then not have any entries and can be dropped. The corresponding indexes can be shifted accordingly. This will keep numerical consistency i.e. worst to best.

5.2. Histograms

The histograms can be found in the accompanying pdf.

6. Action to take

5 main actions will be taken, summarised below;

- -7 Values
  - Change all -7 values to at least the maximum value of that respective feature
- -8 Values
  - See where imputation is possible
  - Where imputation not possible change values “null”.
- Logical integrity
  - Rows failing the logical test will need to be dropped
- Categorical Features
  - Map both numerical scales to a common scale
  - Change all “unknown delinquency” values to “current or never delinquent”
- Outliers
  - Review outliers, checking for validity

7. References

[1] FICO description of data

https://community.fico.com/s/explainable-machine-learning-challenge?tabset-3158a=2

8. Appendix

8.1. Terminology & Assumptions

- “trade” is equivalent to an account/contract with a lender
- “Satisfactory trade” means a trade that has not had a late payment
- “Derogatory comment” is equivalent to a trade that has been defaulted on
- “Utilization” details the fraction of available credit currently used
- “Revolving trade” is equivalent to Credit Account, Credit card in a bank i.e. the credit revolves
- “Instalment trade” is equivalent to a loan with fixed payment instalments
- “Inquiry” is an inquiry made by lender on the applicant’s credit history
- “Delinquency” is a late a payment, either 30, 60, 90 days late
- “External risk estimate” is assumed to be a risk factor applied to the applicant based on their personal economic situation

8.2. Special Values

-9 Values

According to the documentation provided by FICO [1], a value of -9 indicates a “Special Value”. No clear explanation is given. It is likely that no information was available and a value of zero was avoided so the data would not be affected since a value of zero can be a positive or negative marker.

-8 Values

According to the documentation provided by FICO [1], a value of -8 indicates “No Usable/Valid Trades or Inquiries”. It is not stated the reasoning why this data could not be collected. This value appears in 9 features. One option to deal with these would be to use linear regression

to predict what the value would have been based on the most similar other users. The count for “NetFractionInstallBurden” is quite high at approx. 30% of overall entries.

-7 Values

According to the documentation provided by FICO [1], a value of -7 indicates “Condition not Met”. This value appears in 2 features that track the total number of months since a delinquency, “MSinceMostRecentDelq” and an enquiry, “MSinceMostRecentInqexcl7days”. These features are both monotonically decreasing meaning a higher value must result in a lower probability of a bad result. In this case “Condition not met” is a good result and should receive an appropriately high value. Keeping a negative value will distort the results. Since this value appears in 2 features it is not clear what the mapping should be. I suggest it should be a minimum of the max of either feature to begin with.

8.3. Continuous Features

Descriptive Statistics

Feature	count	mean	std	min	25%	50%	75%	max
ExternalRiskEstimate	944	72.05297	9.767572	42	65	72	80	93
MSinceOldestTradeOpen	944	193.4269	100.5874	-8	132	180	252.25	598
MSinceMostRecentTradeOpen	944	9.644068	10.73554	0	3	6	12	97
AverageMInFile	944	77.36653	32.08778	4	56	75	94.25	240
NumSatisfactoryTrades	944	21.1303	11.25728	1	13	20	28	78
NumTrades60Ever2DerogPubRec	944	0.544492	1.100531	0	0	0	1	10
NumTrades90Ever2DerogPubRec	944	0.356992	0.796795	0	0	0	0	6
PercentTradesNeverDelq	944	92.58898	11.0401	33	89.75	97	100	100
MSinceMostRecentDelq	944	8.162076	20.71777	-8	-7	1	17	83
NumTotalTrades	944	22.77542	12.84735	0	14	21	29	100
NumTradesOpeninLast12M	944	1.827331	1.845247	0	0	1	3	19
PercentInstallTrades	944	34.83686	16.87773	0	23	33	45.25	100
MSinceMostRecentInqexcl7days	944	0.074153	5.689641	-8	0	0	1	24
NumInqLast6M	944	1.561441	2.000646	0	0	1	2	16
NumInqLast6Mexcl7days	944	1.485169	1.950553	0	0	1	2	16
NetFractionRevolvingBurden	944	35.2161	29.10794	-8	9	30	57.25	120
NetFractionInstallBurden	944	42.40572	40.96893	-8	-8	55	79	144
NumRevolvingTradesWBalance	944	3.939619	3.314155	-8	2	3	5	21
NumInstallTradesWBalance	944	1.684322	3.396388	-8	1	2	3	14
NumBank2NatlTradesWHighUtilization	944	0.564619	2.579398	-8	0	0.5	2	12
PercentTradesWBalance	944	66.52436	22.16561	-8	50	67	83	100

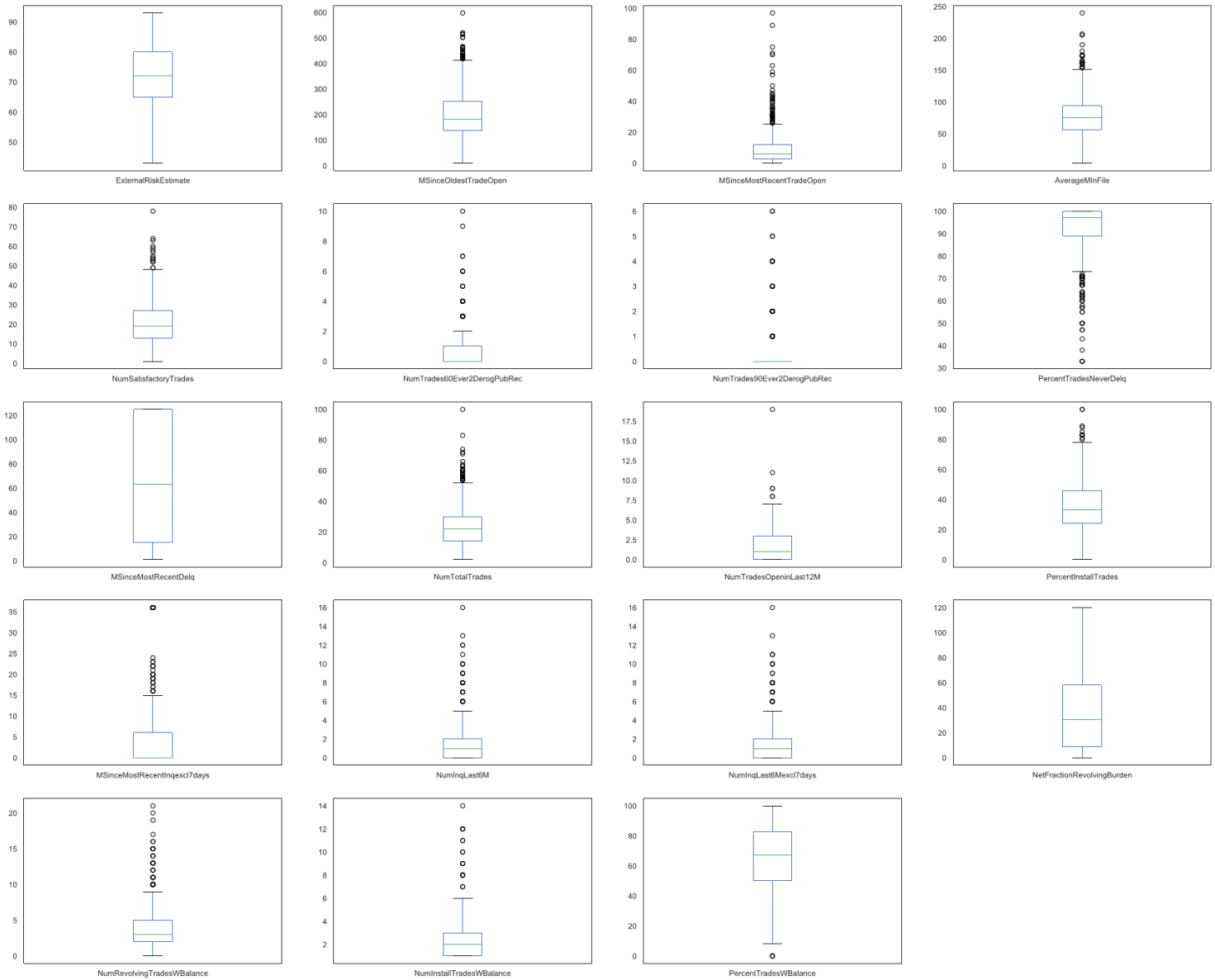
8.4. Categorical Features

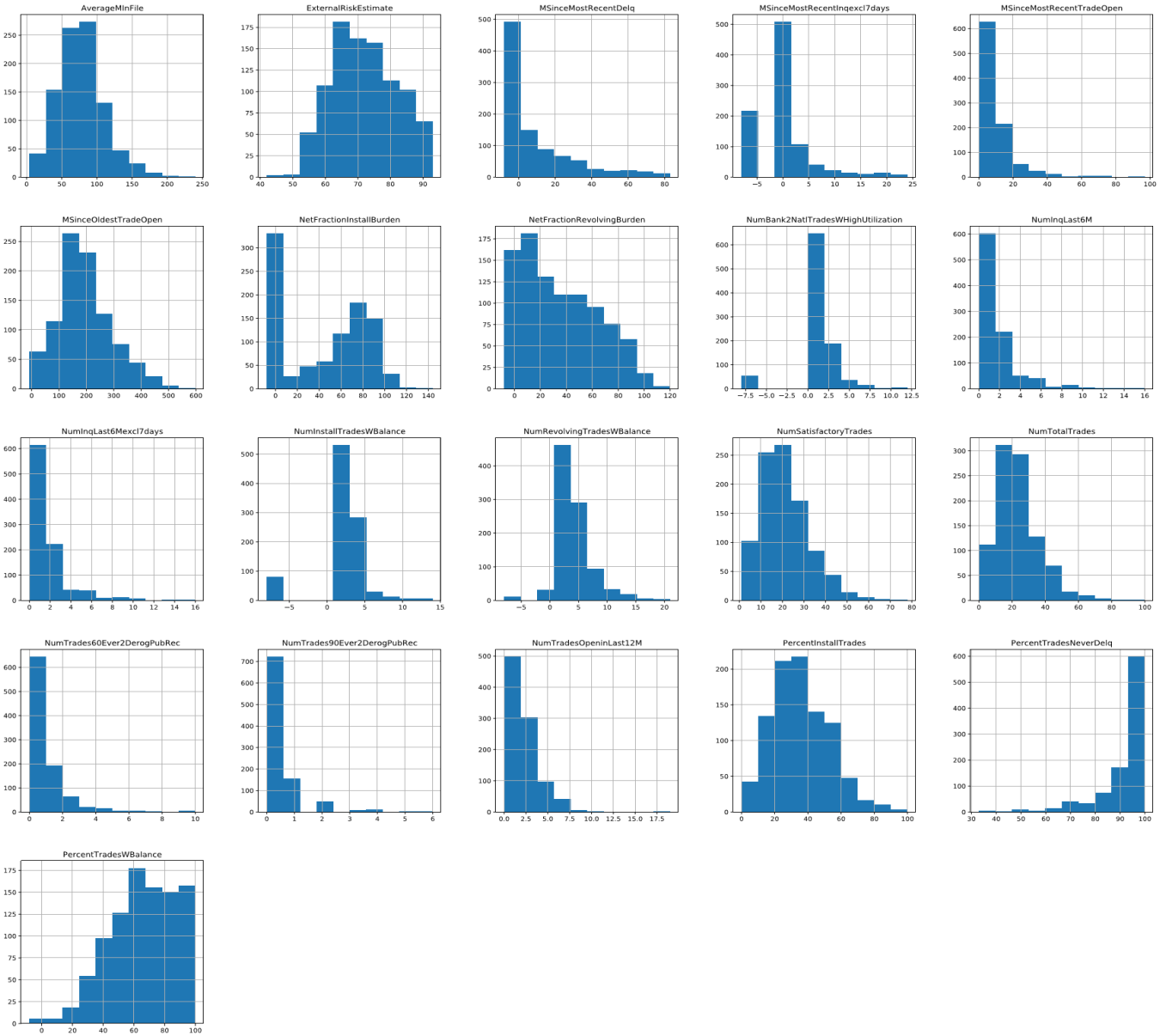
Descriptive Statistics

Column1	count	unique	top	freq
Risk Performance	944	2	Bad	516
MaxDelq2PublicRecLast12M	944	8	7	411
MaxDelqEver	944	7	8	432

8.5. Box Plots & Histograms

See below summary of box plots and histograms. Accompanying pdfs will show larger plots.


ExternalRiskEstimate
MSinceOldestTradeOpen
MSinceMostRecentTradeOpen
AverageMInFile
NumSatisfactoryTrades
NumTrades60Ever2DerogPubRec
NumTrades90Ever2DerogPubRec
PercentTradesNeverDelq
MSinceMostRecentDelq
NumTotalTrades
NumTradesOpeninLast12M
PercentInstallTrades
MSinceMostRecentInqexcl7days
NumInqLast6M
NumInqLast6Mexcl7days
NetFractionRevolvingBurden
NumRevolvingTradesWBalance
NumInstallTradesWBalance
PercentTradesWBalance


AverageMInFile
ExternalRiskEstimate
MSinceMostRecentDelq
MSinceMostRecentInqexcl7days
MSinceMostRecentTradeOpen
MSinceOldestTradeOpen
NetFractionInstallBurden
NetFractionRevolvingBurden
NumBank2NatlTradesWHighUtilization
NumInqLast6M
NumInqLast6Mexcl7days
NumInstallTradesWBalance
NumRevolvingTradesWBalance
NumSatisfactoryTrades
NumTotalTrades
NumTrades60Ever2DerogPubRec
NumTrades90Ever2DerogPubRec
NumTradesOpeninLast12M
PercentInstallTrades
PercentTradesNeverDelq
PercentTradesWBalance