Data Quality Report

The data quality report uses tables and visualisations to explore the data. For each feature in the dataframe, we prepare summary tables (with descriptive stats) and plots.

General considerations

The dataset represents HELOC applications made by real homeowners. A HELOC is a line of credit typically offered by a bank as a percentage of home equity.
More details about the features and the composition of the dataset can be found in the dictionary and on https://community.fico.com/s/explainable-machine-learning-challenge?tabset-3158a=4fbc8.

In summary

- The dataset is purposely split roughly 50/50 between ‘Good’ and ‘Bad’ target outcomes
- Data has been given special values:
    - -9 loan applicant’s credit bureau report not investigated or not found
    - -8 for no valid/usable information
    - -7 for no information of that type (e.g. no inquiries, no delinquencies)
- what should be a slash “/” in the variable names got replaced with a “2”
- definitions:
    - Trade: Every credit agreement between the consumer and a lending institution
    - Inquiry: when a lending institution has pulled a consumer’s credit bureau report in order to make a credit decision
    - delinquency: payment received some period of time past its due date. This is typically measured in 30-day intervals

Features meaning

MSinceMostRecentDelq: Months Since Most Recent Delinquency.

MSinceMostRecentInqexcl7days: Months Since Most Recent Inquiry excluding 7days. Excluding the last 7 days removes inquiries that are likely due to price comparison shopping.

ExternalRiskEstimate: Consolidated version of risk markers. The feature represents a flag for the target outcome.

MSinceOldestTradeOpen: Months Since Oldest Trade Open.

MSinceMostRecentTradeOpen: Months Since Most Recent Trade Open.

AverageMInFile: Average Months in File.

NumSatisfactoryTrades: Number Satisfactory Trades.

NumTrades60Ever/DerogPubRec: Number Trades 60+ Ever. A "satisfactory trade" is one where the borrower has paid on time as agreed.

NumTrades90Ever/DerogPubRec: Number Trades 90+ Ever.

PercentTradesNeverDelq: Percent Trades Never Delinquent.

NumTotalTrades: Number of Total Trades (total number of credit accounts).

NumTradesOpeninLast12M: Number of Trades Open in Last 12 Months.

PercentInstallTrades: Percent Installment Trades. Installment trade accounts involve agreements you make to pay an account over time. These accounts show your original and current balance on your credit report, as well as the amount you’re required to pay each month. Unless an installment account is new, your current balance should be less than the original balance as long as you’re making payments on time. Examples of common installment accounts include auto loans, mortgages and personal loans from banks or finance companies. (source: https://budgeting.thenest.com/open-trades-credit-report-23674.html)

NumInqLast6M: Number of Inquiries Last 6 Months.

NumInqLast6Mexcl7days: Number of Inq Last 6 Months excluding 7days. Excluding the last 7 days removes inquiries that are likely due to price comparison shopping.

NetFractionRevolvingBurden: Net Fraction Revolving Burden. This is revolving balance divided by credit limit. A revolving balance is the portion of credit card spending that goes unpaid at the end of a billing cycle.

NetFractionInstallBurden: Net Fraction Installment Burden. This is installment balance divided by original loan amount.

NumRevolvingTradesWBalance: Number Revolving Trades with Balance. Revolving trade lines are credit products that creditors can use multiple times. These accounts include credit cards and equity lines. The accounts "revolve," meaning the balances fluctuate from month to month based on usage. The term "trade" simply means account. The balance you owe, relative to the maximum line amount, has an impact on your overall credit score. (source: https://www.sapling.com/7839565/do-lines-mean-credit-bureau).

NumInstallTradesWBalance: Number Installment Trades with Balance. Installment trade accounts involve agreements you make to pay an account over time.

NumBank/NatlTradesWHighUtilization: Number Bank/Natl Trades with high utilization ratio. This counts the number of credit cards on a consumer credit bureau report carrying a balance that is at 75% of its limit or greater.

PercentTradesWBalance: Percent Trades with Balance.

The dataset

The initial dataframe is composed of 1000 rows and 24 columns with no apparent missing data. Exploring the datatypes, the first attribute that can be classified as categorical is ***RiskPerformance*** which represents a flag for the target outcome. The file 'data_dictionary', sheet 'MaxDelq', shows that the values for the attributes ***MaxDelq***/***PublicRecLast12M*** and ***MaxDelqEver*** refer to specific categories. The values were therefore replaced with the corresponding category and the attributes were labelled as categorical.
Two additional feature are treated as categorical: ***MSinceMostRecentDelq*** and ***MSinceMostRecentInqexcl7days***. Even if the value of the feature represents a certain number of months, these two features contain a substantial amount of special values -7 which indicates no inquiries or no delinquencies. Therefore, in order to keep this information, which may be useful to see certain pattern in the target outcome, the feature is converted into categorical where the values originally assigned to -7 are replaced with the explicit meaning of the value and the numerical data is grouped in different ranges of values using equal size bins. In more details, the feature ***MSinceMostRecentDelq***, which represents the number of months that have passed since the most recent delinquency was recorded, the original values have been placed in equal size bins of ten months' range. The feature ***MSinceMostRecentInqexcl7days***, which represents the number of months that have passed since the most recent inquiry was recorded (excluding 7 days), as been treated with the same approach with the only difference that the size of the bins is of 5 months (this decision was based on the fact that the range of original values was considerably smaller).

We adopted this approach over others that were considered and are listed below:

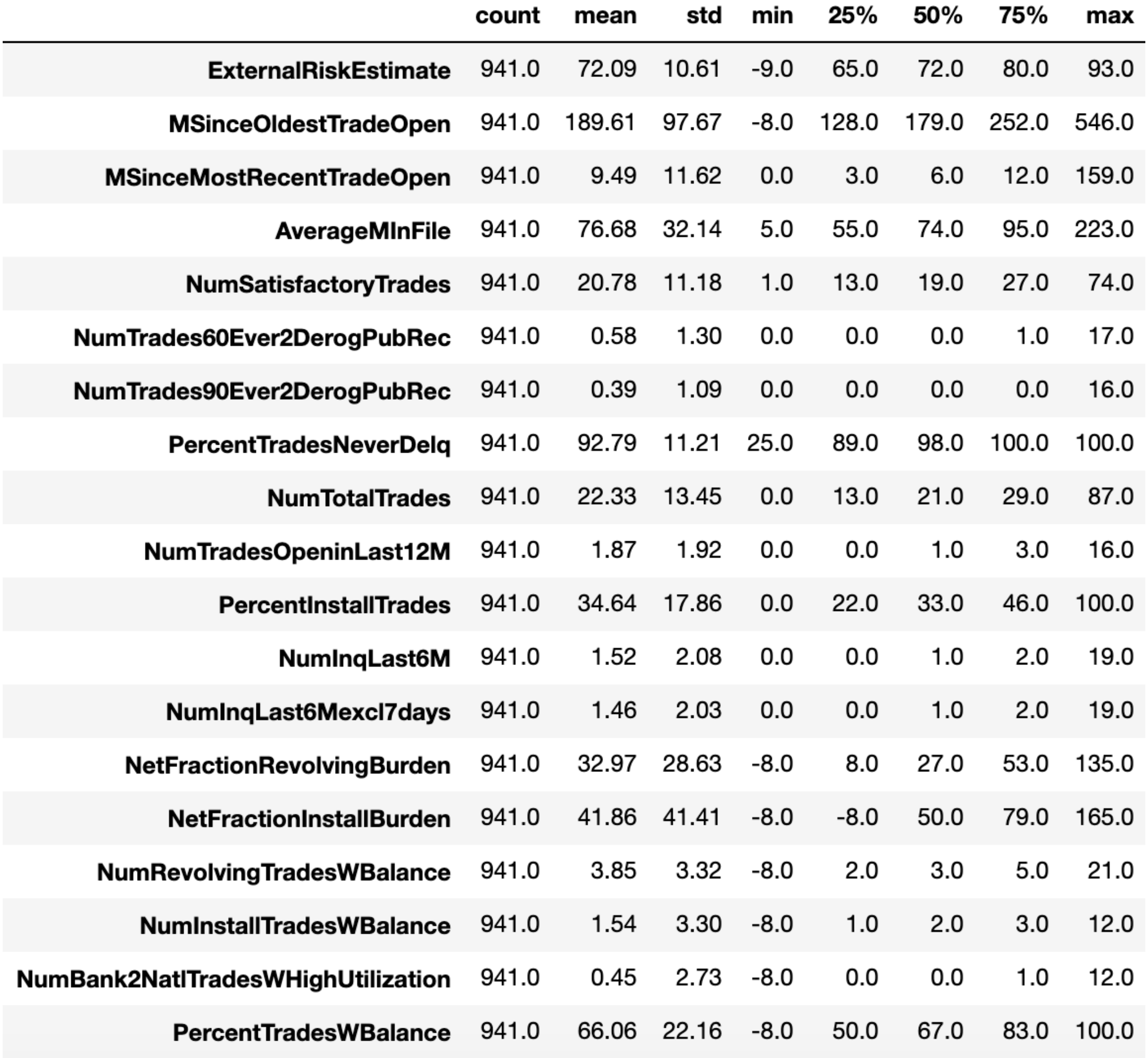
- keep features as numerical: in this scenario the special values -7 should have been replaced by a meaningful value. But this is not an easy decision to make because, for example, assigning really high values (which would mean that the delinquency happened a long time ago) would have an impact on all the stats of the features.
- Drop the feature: another option considered was to drop the feature ***MSinceMostRecentDelq***, affected by 50% of special values (-7 and -8), but this would result in an important data loss as the special value -7 holds a special meaning of 'no delinquencies' which could possibly be link to the target outcome.
- Replace values with imputation: this option was considered for the feature ***MSinceMostRecentInqexcl7days*** where about 20% of values was labelled as -7. However, replacing the special values with a feature stat (e.g. mean, median) would have essentially modified the real meaning of the special value, 'no inquiries', which could possibly be link to the target outcome.

The dataframe does not contain any duplicate columns but contains duplicate rows. The analysis shows that the duplicate rows are highlighted as such since they all show the same value, -9, for every column. From the FICO Community challenge webpage (https://community.fico.com/s/explainable-machine-learning-challenge?tabset-3158a=2) and the file 'data_dictionary', sheet 'SpecialValues', the value indicates no record which means no credit history/score information is available. Since the rows containing only -9 values don't hold any information, these were dropped.

The dataframe does not contain any constant column i.e. a column with cardinality one, therefore no column was dropped.
The update/cleaned dataframe was saved as:
'CreditRisk_clean_round1_25Feb2019_DataQualityReport.csv'

Tables

Continuous Features statistics table



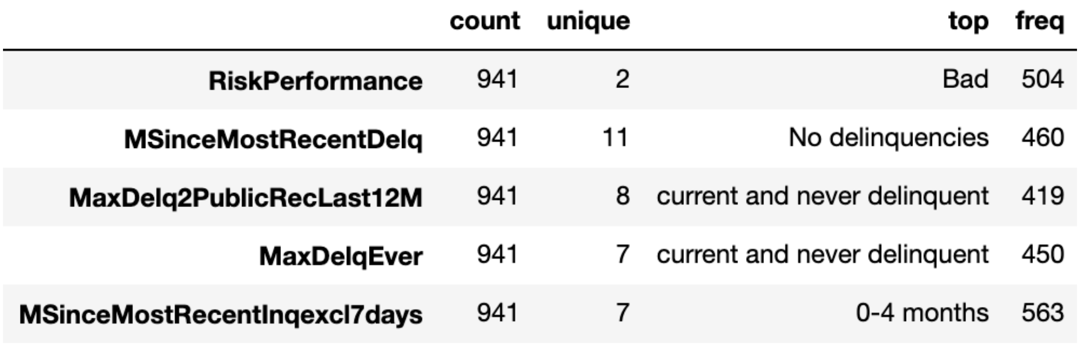
	count	mean	std	min	25%	50%	75%	max
ExternalRiskEstimate	941.0	72.09	10.61	-9.0	65.0	72.0	80.0	93.0
MSinceOldestTradeOpen	941.0	189.61	97.67	-8.0	128.0	179.0	252.0	546.0
MSinceMostRecentTradeOpen	941.0	9.49	11.62	0.0	3.0	6.0	12.0	159.0
AverageMInFile	941.0	76.68	32.14	5.0	55.0	74.0	95.0	223.0
NumSatisfactoryTrades	941.0	20.78	11.18	1.0	13.0	19.0	27.0	74.0
NumTrades60Ever2DerogPubRec	941.0	0.58	1.30	0.0	0.0	0.0	1.0	17.0
NumTrades90Ever2DerogPubRec	941.0	0.39	1.09	0.0	0.0	0.0	0.0	16.0
PercentTradesNeverDelq	941.0	92.79	11.21	25.0	89.0	98.0	100.0	100.0
NumTotalTrades	941.0	22.33	13.45	0.0	13.0	21.0	29.0	87.0
NumTradesOpeninLast12M	941.0	1.87	1.92	0.0	0.0	1.0	3.0	16.0
PercentInstallTrades	941.0	34.64	17.86	0.0	22.0	33.0	46.0	100.0
NumInqLast6M	941.0	1.52	2.08	0.0	0.0	1.0	2.0	19.0
NumInqLast6Mexcl7days	941.0	1.46	2.03	0.0	0.0	1.0	2.0	19.0
NetFractionRevolvingBurden	941.0	32.97	28.63	-8.0	8.0	27.0	53.0	135.0
NetFractionInstallBurden	941.0	41.86	41.41	-8.0	-8.0	50.0	79.0	165.0
NumRevolvingTradesWBalance	941.0	3.85	3.32	-8.0	2.0	3.0	5.0	21.0
NumInstallTradesWBalance	941.0	1.54	3.30	-8.0	1.0	2.0	3.0	12.0
NumBank2NatlTradesWHighUtilization	941.0	0.45	2.73	-8.0	0.0	0.0	1.0	12.0
PercentTradesWBalance	941.0	66.06	22.16	-8.0	50.0	67.0	83.0	100.0

The main issue that can be identified from the table is in the negative values found in the min (and 25%) for several attributes (for example: ***ExternalRiskEstimate***, ***MSinceOldestTradeOpen***, ***NetFractionInstallBurden*** and others). It's important to remember that these negative values have different meanings and should be handled accordingly.

It can be observed that the count for every attribute is equal which shows that each attribute doesn't contain any null values.

Another possible issue can be observed in the max attribute for the features where this value is much greater than the mean value (for example: ***MSinceOldestTradeOpen***, ***NumTrades60Ever***/***DerogPubRec***, ***NumBank***/***NatlTradesWHighUtilization*** and others).

Categorical Features statistics table



	count	unique	top	freq
RiskPerformance	941	2	Bad	504
MSinceMostRecentDelq	941	11	No delinquencies	460
MaxDelq2PublicRecLast12M	941	8	current and never delinquent	419
MaxDelqEver	941	7	current and never delinquent	450
MSinceMostRecentInqexcl7days	941	7	0-4 months	563

The table shows that the feature ‘***RiskPerformance’*** has more ‘Bad’ than ‘Good’ but the 2 values are balanced. FAQ 1 on the FICO Community challenge webpage (https://community.fico.com/s/explainable-machine-learning-challenge?tabset-3158a=4fbc8) explains that the dataset was purposely populated with about 50/50 between ‘Good’ and ‘Bad’.

Both the features ***MaxDelq***/***PublicRecordsLast12M*** and ***MaxDelqEver*** appear to have, for about half their dataset, a value of ‘current and never delinquent’.

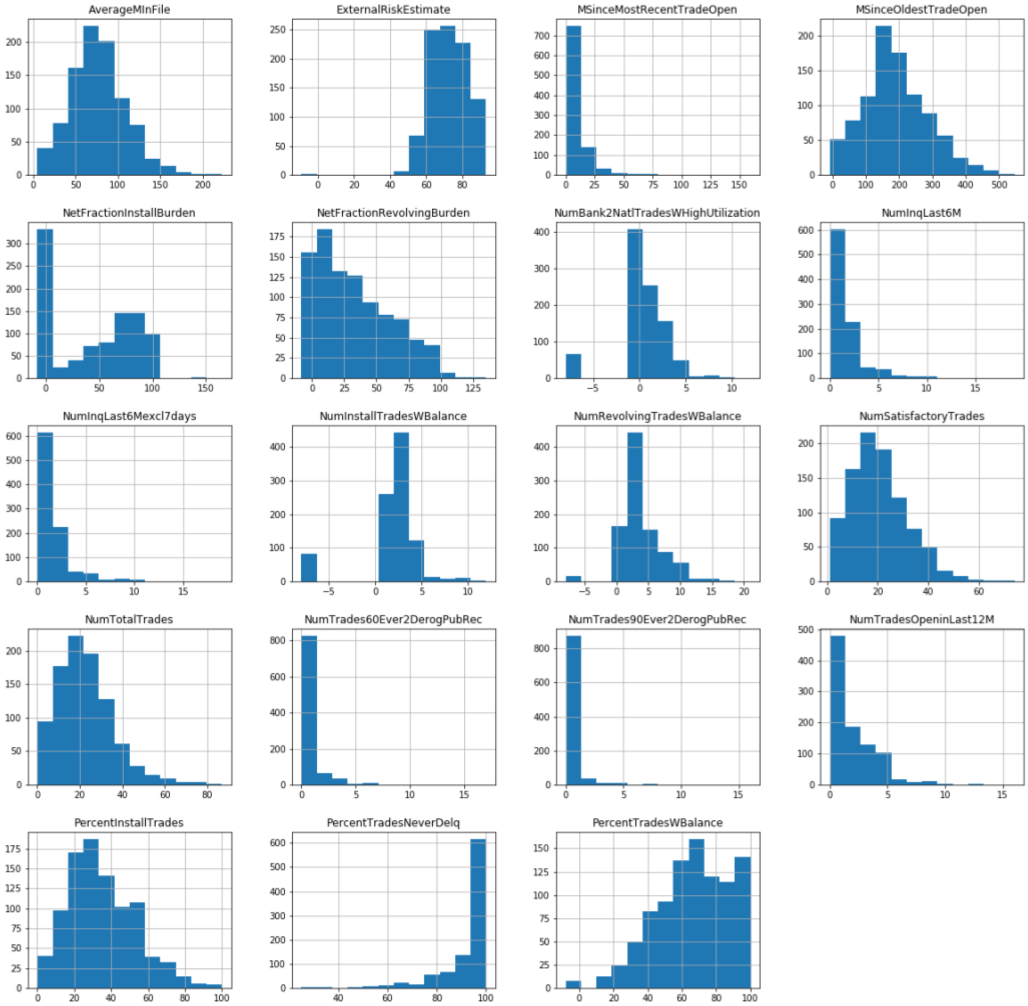
The feature ‘***MSinceMostRecentDelq’*** shows that the majority of values, about 50%, shows no delinquencies while the feature ‘***MSinceMostRecentIneescl7days’*** shows that more than 50% of the instances had an inquiry in the last months.

The table shows than no features have a unique value, i.e. cardinality 1.

Plots

Continuous Features: histograms

The figure below contains the histograms for every continuous feature



A few considerations of what we can understand from these plots:

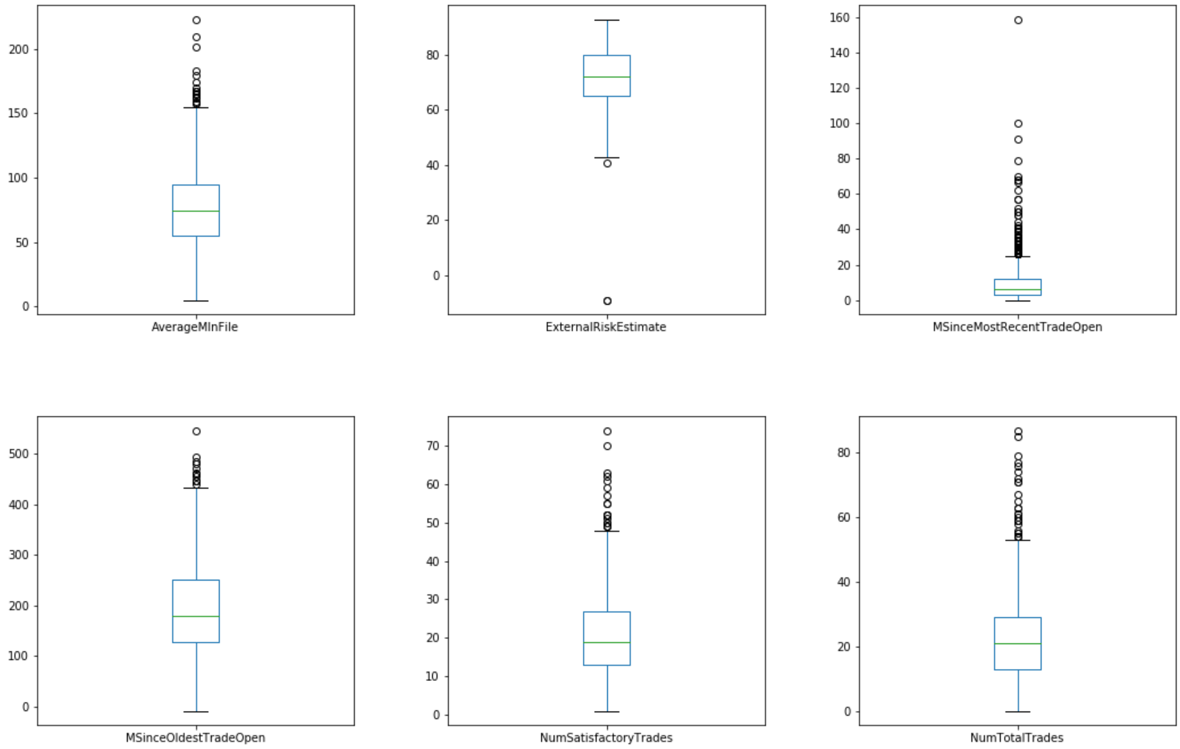
- certain features show a ‘Gaussian’ or ‘bell’ shape which indicates a normal distribution (e.g. ***AverageMinFile***, ***ExternalRiskEstimate***, ***MSinceOldestTradeOpen***, ***PercentInstallTrades***)
- other features show a unimodal (skewed) distribution towards right, for example ***NetFractionRevolvingBurden***, ***NumSatisfacotoryTrades***, ***NumTotalTrades***. And towards left, like ***PercentTradesWBalance***
- some features show an exponential distribution with many values near 0, for example ***MSinceMostRecentTradeOpen***, ***NumInqLast6M***, ***NumInqLast6Mexcl7days***,

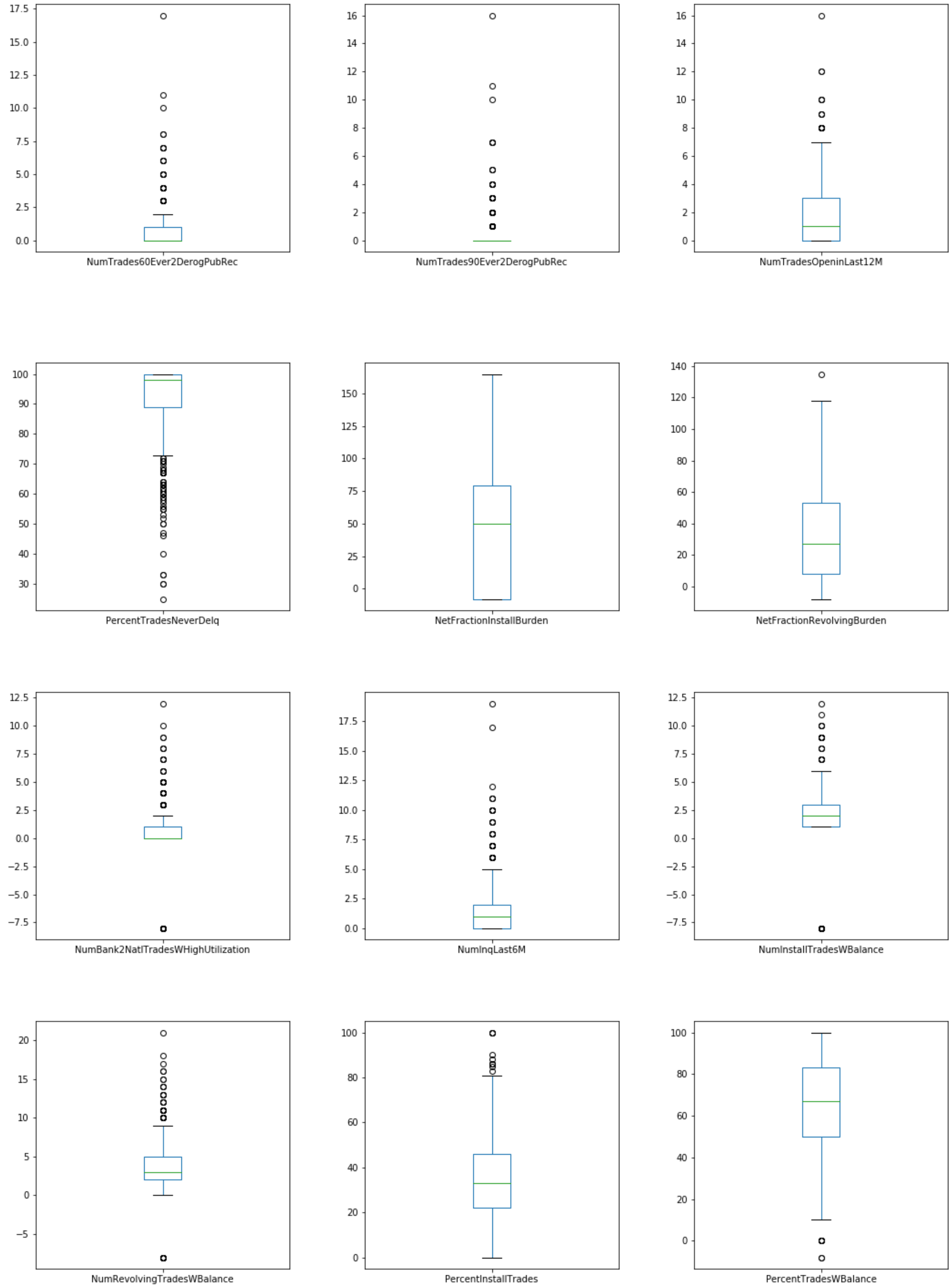
NumTrades60Ever*/*DerogPubRec, ***NumTrades90Ever*/*DerogPubRec***, ***NumTradesOpeninLast12M*** or with high values near 100% as in ***PercentTradesNeverDelq***

- The features not listed above show various distribution shapes but are all also showing some issues related to the fact that a certain amount of data has values below 0, we should remember that the values below zero have a special meaning which refer to missing or unavailable data or are applied when a certain condition is not met. For some features the number of values below zero is quite low, for example ***NumRevolvingTradesWBalance***, ***NumInstallTradesWBalance*** and ***NumBank*/*NatlTradesWHighUtilization*** where their main shape can be recognized. However, the feature ***NetFractionInstallBurden*** shows a high number of negative values which indicates some issues in the data and should be analysed in more detail in order to establish the best approach in handling the feature.

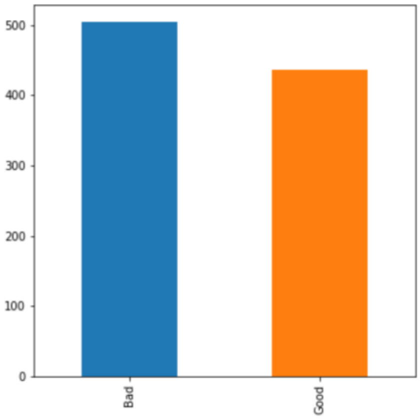
Continuous Features: box plots

The boxplot figures below give a better idea of the range of the values and the stats of the features and also help understand the issues with the negative values discussed above. These plots also point out that almost every feature has outliers and, while in some these are quite limited (e.g. ***NetFractionRevolvingBurden***, ***ExternalRiskEstimate***) in some features these are quite consistent (e.g. ***MSinceMostRecentTradeOpen***, ***PercentTradesNeverDelq***, ***NumTotalTrades***, etc…). Outliers are therefore another issue in the data that need to be addressed.

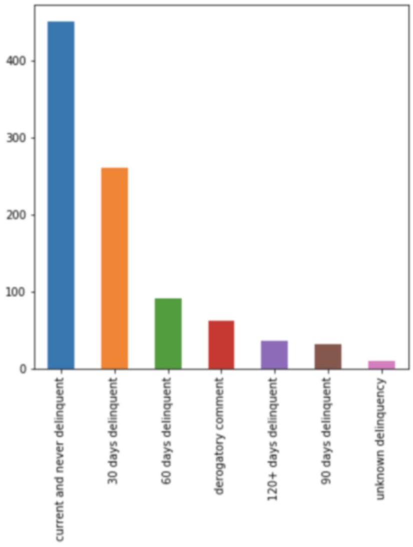



17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
NumTrades60Ever2DerogPubRec
16
14
12
10
8
6
4
2
0
NumTrades90Ever2DerogPubRec
16
14
12
10
8
6
4
2
0
NumTradesOpeninLast12M
100
90
80
70
60
50
40
30
PercentTradesNeverDelq
150
125
100
75
50
25
0
NetFractionInstallBurden
140
120
100
80
60
40
20
0
NetFractionRevolvingBurden
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
NumBank2NatlTradesWHighUtilization
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
NumInqLast6M
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
NumInstallTradesWBalance
20
15
10
5
0
−5
NumRevolvingTradesWBalance
100
80
60
40
20
0
PercentInstallTrades
100
80
60
40
20
0
PercentTradesWBalance

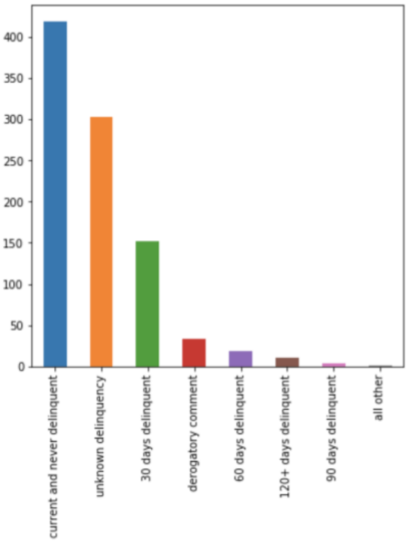
Categorical Features: bar plots



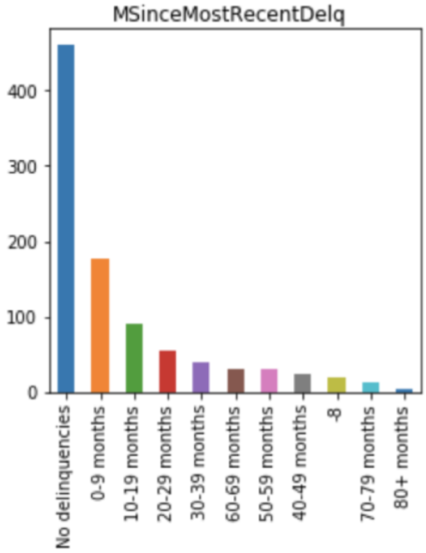
RiskPerformance

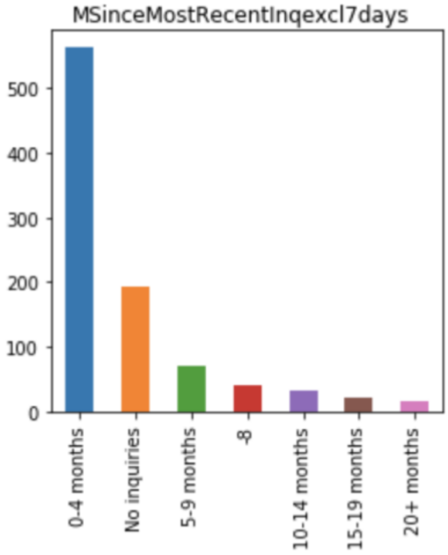


MaxDelqEver



MaxDelq/PublicRecLast12M





The bar plots give us more information about the categorical features such as:

- The ***RiskPerformance*** feature, which represent the target value, shows an almost 50/50 distribution. We should remember that the dataset has been created on purpose with a 50/50 proportion of ‘Good’ and ‘Bad’.
- Both ***MaxDelqEver*** and ***MaxDelq/PublicRecLast12M*** have as the most common value ‘current and never delinquent’ and then a rapid decrease towards different delinquency categories. A difference worth mentioning is that the feature ***MaxDelq/PublicRecLast12M*** has a high value of ‘unknown delinquency’, probably because the data refers to the last 12 months, while in ***MaxDelqEver*** this category is very low. Another common trend that can be observed is that the delinquency decreases as the days of overdue payment increase, this shows that only a few instances were very late in their payments (90 and 120+ days) but a larger amount of instances were late in their payments for 30 and 60 days.
- ***MSinceMostRecentDelq*** has most of the values as ‘no delinquencies’ with a rapid decrease towards increasing months since the last delinquency event. This means that most of the

instances have no recorded delinquencies but, on the other hand, the majority of the instances with recorded delinquencies refer to events that happened recently rather than years ago.

- ***MSinceMostRecentInqexcl7days*** has most of the values as 0-4 months meaning that most of the instances had an inquiry recently. The second most common value is related to no inquiries and only a minor part of instances has had inquires more than 5 months ago.
- Finally, we can observe that there are no negative values i.e. -9, corresponding to missing values. However, the features ***MSinceMostRecentDelq*** and ***MSinceMostRecentInqexcl7days*** show a small presence of special values -8 which will be analysed in more detail in the next steps.