

email: [sebastian.bloy.stats@gmail.com](mailto:sebastian.bloy.stats@gmail.com)

Linkedin: <https://www.linkedin.com/in/sebastian-bloy-ab1741178>

R-code available at: [https://github.com/sebastianbloy80/bondora\\_credit\\_default](https://github.com/sebastianbloy80/bondora_credit_default)

## Illustrative statistical analysis of credit defaulting probability using Bondora raw data

---

### Last Change:

09/25/20: Finalizing of document

## Content

---

Illustrative statistical analysis of credit defaulting probability using Bondora raw data.....	1
Disclaimer.....	2
Info on use of colors - Color Blindness Info.....	3
1.Motivation.....	4
1.1 Target and motivation of the analysis.....	4
1.2 Origin of Data.....	4
2.Raw Data – Origin, Software used for analysis and Data type overview.....	5
2.1 Raw Data – Origin.....	5
2.2 Used Software and libraries.....	5
2.3 Data Type Overview.....	5
3.Define relevant Data, Feature Calculation, Data Cleaning, Outlier Detection and Test of significance.....	7
3.1 Defining the singular dependent and multiple independent Variables – a priori.....	7
3.1.1 Defining the dependent Variable.....	8
3.1.2 Defining the independent Variables – a priori.....	8
3.2 Feature Calculation.....	9
3.3 Creation of Loan Analysis Dataset – Filtering out further insignificant data.....	10
3.4 Data cleaning, presumed Errors and outlier detection.....	10
3.4.1 Missing data for multiple variables after 2017.....	10
3.4.2 Treatment of outliers.....	12
3.4.3 Strange Correlations between Amount and ManualBids for data before October 2012.....	12
3.4.4 Monthly Payment always zero or missing until mid 2014.....	14
3.4.5 Deletion of low count Levels of Factor.....	14
4.Building the Logistic Model.....	15
4.1 Test of significance.....	15
4.2 Detection of collinearity – and its avoidance.....	16
4.3 Running a simple glm – how significant are the independent variables.....	17
4.4 Calculating McFaddens Pseudo-R2 – Stepwise addition of independent variables and testing prediction capability.....	19
4.5 Graphical depiction of predictions with model #5 and #7.....	21
5.Final note on the results.....	23
6.Appendix.....	24
6.1 Raw Data – Data Types after initial import.....	24
6.2 Raw Data – Graphical overview of independent variables [not cleaned yet].....	26
6.3 Raw Data – Table of Correlation between BidsManual and Amount.....	29
6.4 Analysis Data – Graphical overview of independent variables [cleaned].....	31

---

## Disclaimer

---

The information contained in this document and possible resources available for download are not intended as, and shall not be understood or construed as, financial advice. I am not an attorney, accountant or financial advisor, nor am I holding myself out to be. This document and the information contained within does not substitute financial advice from professionals.

The data used in this analysis has been taken from an external source. I cannot and will not guarantee the correctness of said data, nor can or will I guarantee the correctness of any conclusion drawn from it. Furthermore all used statistical methods must not represent the optimal course of analysis and may be subject to later changes.

Use all information given in this document at your own discretion and risk.

**Info on use of colors - Color Blindness Info**

---

Color blind friendly color palettes are used within this document where regarded as necessary. You will find the usage of red, green, blue etc throughout this document in combination with methods that enable color blind readers to distinguish one color from another by usage of a combination of symbols or dotted/dashed lines to support visual distinction. Where not possible a color blind friendly color palette will be used.

Please excuse, if I do not provide easily distinguishable combinations all the time as I am not affected by color blindness but try to provide you with a good enough combination.

## 1. Motivation

---

### 1.1 Target and motivation of the analysis

It is the target of this analysis to try and find a statistical model from data provided by Bondora itself which enables the prediction of the probability of a loan to be defaulting or not defaulting for the entirety of its duration. This prediction shall result in a binary outcome. The analysis is motivated by pure personal interest in the matter.

### 1.2 Origin of Data

Bondora Capital OÜ<sup>1</sup> is providing a web-plattform for so called peer-to-peer creditory services. Private investors can buy into credits taken by private persons from Estonia, Finland, Slovakia and Spain. The credit itself is given out by a Bank. Bondora provides the platform to partake in those credits as a private investor.

The whole credit itself is not part of the offer but it is split up into several piece of differing amount.

If the investor decides to invest into loans by his own selection, Bondora itself provides a credit grading system, ranging from „AA“ (safest grade) up to „HR“ (highest risk). There is also a „Credit Score“ Value which ranging from „0“ to „1000“ whereas the higher the number indicates a more positive payment behavior of the borrower than can be assumed with lower numbers<sup>2</sup>.

This analysis does not try to criticize Bondora's credit grading system in any way. It merely uses the data kindly provided on Bondora's web-page freely to be analyzed.

---

<sup>1</sup> A.H.Tammsaare tee 47, Tallinn 11316, Estonia, Estonian company number 12831506, VAT number EE101252401

<sup>2</sup> A more detailed list of definition can be found under: <https://www.bondora.com/en/public-reports>

## 2. Raw Data – Origin, Software used for analysis and Data type overview

In this chapter I will show what the origin of the raw data is, how it has been cleaned up and what variables have been identified as relevant for the logistic regression.

### 2.1 Raw Data – Origin

The raw data used here has been taken from the Website of Bondora itself. Under the web-address <https://www.bondora.com/en/public-reports> a csv-file named „LoanData“ is encapsulated inside a zipped file and must be unzipped for further use. The Download link is named “Loan dataset” at the website.

Each row (except for the first one) represents an individual loan which is described by the 112 variables organized in columns. It is important to note, that these are already assigned loans – meaning there is a p2p-lender on the loan itself. These are not new loans but already assigned ones.

Name	Format	Rows	Columns	ReportAsOfEOD	Size (unzipped)
LoanData	*.csv	150623	112	07.09.20	122.803 KB

The data file was imported to R for further analysis.

### 2.2 Used Software and libraries

The following Software has been used for analysis.

Software	Version	Platform
R	3.6.1 (2019-07-05) -- "Action of the Toes"	x86_64-w64-mingw32/x64 (64-bit)

The following libraries have been used for analysis (Dependencies not shown).

Library	Version
pscl	01.05.05
readr	01.03.01
dplyr	0.8.0.1
DMwR	0.4.1
zoo	1.8-5
corrplot	0.84
car	3.0-9
Caret	6.0-83
MASS	7.3-51.3
reshape	0.8.8

### 2.3 Data Type Overview

After successful import of the csv-file, a quick glance at the data types have been done. The data types are according to the given R-typification logic<sup>1</sup>.

As can be seen most of the data is numerical or of the date/time type. Some are identified as character. This all seems to be an “artifact” of the importing function used (read\_csv()) as some of them surely should be defined as a factor. R’s read\_csv() did a good job in important most of them right, especially time related variables. What kind of type is really to be awaited in the raw data – this will be done manually after the import itself by eyeballing and re-coding them if deemed necessary.

<sup>1</sup> A complete overview of all data types after import can be found in the Appendix section of this writing

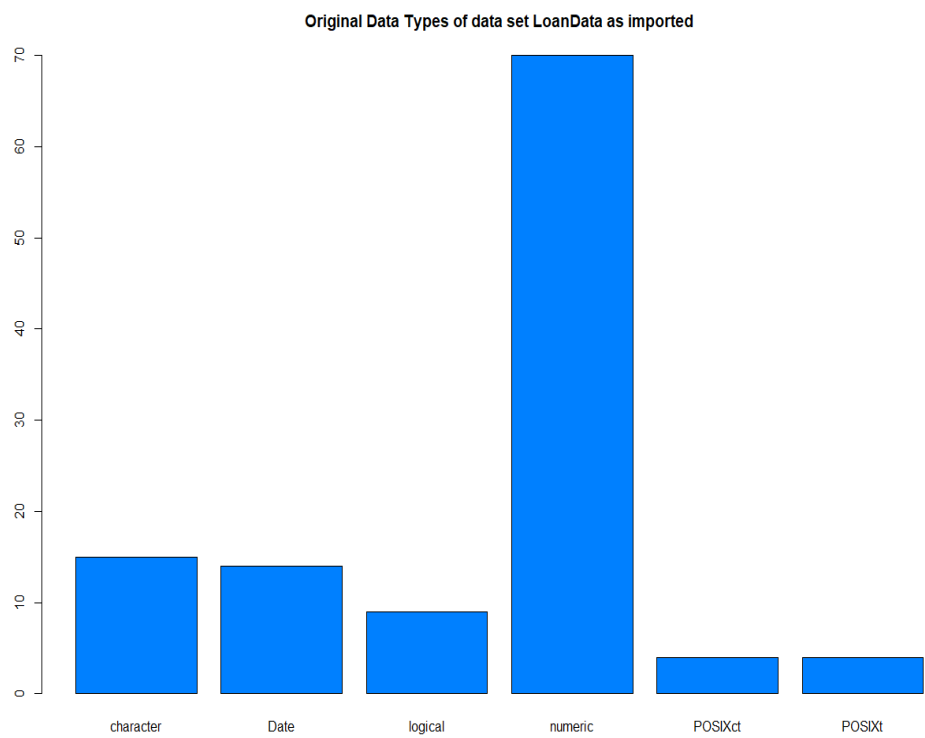


Figure 1: Types of variable Data types - identified by R on import of Loan Data set

### 3. Define relevant Data, Feature Calculation, Data Cleaning, Outlier Detection and Test of significance

---

For the definition of a logistic model to determine if a loan is going to default it is imperative to...

- define the dependent Variable
- define the independent Variable

in the following text I am trying to rationalize how this selection has been done.

#### 3.1 Defining the singular dependent and multiple independent Variables – a priori

The target of this analysis is the ability to provide a statistical model with a rather high chance of correctly predicting if a loan will be defaulting over the course of its lending time. For this we have to define first, what defaulting means in the borders of this analysis. I just take the definition provided by Bondora itself:

*"We have a more conservative approach and consider a loan defaulted when a borrower has missed a scheduled payment by more than 60 days (or, in other words, the loan is 60+ days overdue). We do this in order to indicate that we have started legal action against the borrower, not that the loan has been written off."*<sup>1</sup>

Scheduled payment can be described as a prefixed Day (via the loan contract) of payment per month at which the borrower has to provide the loans principal- and interest-amount to the lender (bank). As described in the quote, if no payment has been done after more than 60 days – Bondora does define the loan as defaulted. So defaulted in this manner does not constitute a full loss of a loan, but the start of legal actions after 60+ days Definition of the dependent Variable. This is also the definition I am following in regarding a loan as being defaulted.

---

<sup>1</sup> <https://www.bondora.com/blog/what-happens-if-loan-defaults/> (Date 09/09/2020)

### 3.1.1 Defining the dependent Variable

The Bondora loan data-set provides multiple Variables, which could be used as an indicator of a defaulted loan. I preferred to use "WorseLateCategory" which provides categorical values of the following definition:

*"Displays the last longest period of days when the loan was in Principal Debt"*<sup>1</sup>

So by respecting the definition of a defaulted Loan by Bondora, we define that a loan has been defaulted at anytime of his existence when any of the following categories for "WorseLateCategory" are true:

"61-90", "91-120", "121-150", "15-180" or "180+"

The name for the calculated variable is chosen as "Defaulted" and can take the values of "0" for not defaulted and "1" for defaulted.

Dependent Variablename	Possible values	Vartype
Defaulted	0;1 (1 == Defaulted)	numeric

### 3.1.2 Defining the independent Variables – a priori

An a priori approach is used to select possible independent files which may have a significant impression on the dependent variable, meaning in the first step data is selected by believe of relevance to the model. The relevance of these selected variables will be of course tested and any variable deleted if deemed necessary because of low to non significance to the model, collinearity or variance inflation.

It must be clear that risk assessing variables that are already presented in the LoanData file will not be used. The scope of this analysis is to calculate the risk of a loan to default by the given variables which do not provide already pre-calculated measures of risk. The Bondora data set delivers such measures but I will not take them into consideration as I want to try to define a risk model of defaulting by non-premeasured variables. Therefore the following variables **will not be implemented into the statistical model of defaulting probability**<sup>2</sup> even if they may improve the statistical model:

ExpectedLoss, LossGivenDefault, ExpectedReturn, ProbabilityOfDefault, PrincipalOverdueBySchedule, PlannedPrincipalPostDefault, PlannedInterestPostDefault, EAD1, EAD2, PrincipalRecovery, InterestRecovery, RecoveryStage, StageActiveSince1, StageActiveSince2, ModelVersion, Rating, EL\_V0, Rating\_V0, EL\_V1, Rating\_V1, Rating\_V2, Restructured, ActiveLateCategory, WorseLateCategory, CreditScoreEsMicroL, CreditScoreEsEquifaxRisk, CreditScoreFiAsiakasTietoRiskGrade, CreditScoreEeMini

However the following variables are regarded as independent variables from the start of this analysis and may be dropped if seen fit:

Independent Variable-name	Possible values	Vartype	Hint
NewCreditCustomer	TRUE/FALSE	logical	
LanguageCode	Multiple levels	factor	
Age	Whole numbers	numeric	
Gender	Multiple levels	factor	
Country	Multiple levels	factor	
AppliedAmount	Decimal numbers	numeric	
FreeCash	Decimal numbers	numeric	
DebtToIncome	Decimal numbers	numeric	
Amount	Decimal numbers	numeric	

<sup>1</sup> <https://www.bondora.com/en/public-reports> (Date 09/09/2020)

<sup>2</sup> Of course it is interesting to see how some of these pre-calculated variables will hold up against the probability of defaulting risk from my model. A comparison of pre-calculated "ProbabilityOfDefault" with the calculated risk of the dependent variable "Defaulted" will give insights to this later on in the document.



Independent Variable-name	Possible values	Vartype	Hint
Interest	Decimal numbers	numeric	
LoanDuration	Whole numbers	numeric	Could also be factorial/categorical as values are fixed and providing the duration of the paying back phase in number of months – I decided to go with numeric
MonthlyPayment	Decimal numbers	numeric	
Education	Multiple levels	factor	
EmploymentDurationCurrentEmployer	Multiple levels	factor	
HomeOwnershipType	Multiple levels	factor	
IncomeTotal	Decimal numbers	numeric	
ExistingLiabilities	Whole numbers	numeric	
LiabilitiesTotal	Decimal numbers	numeric	
RefinanceLiabilities	Whole numbers	numeric	
NoOfPreviousLoansBeforeLoan	Whole numbers	numeric	
AmountOfPreviousLoansBeforeLoan	Decimal numbers	numeric	
PreviousRepaymentsBeforeLoan	Decimal numbers	numeric	
PreviousEarlyRepaymentsBeforeLoan	Decimal numbers	numeric	
PreviousEarlyRepaymentsCountBeforeLoan	Whole numbers	numeric	

### 3.2 Feature Calculation

Besides already existing variables the quality of the predictory statistical model could be improved by calculation of new features. The following features have been calculated – their motivation is:

Independent Feature Variablename	Possible values	Vartype	Calculation; Motivation
NeedvsHave	TRUE/FALSE	logical	AppliedAmount – Amount; Could be an indicator of needs the borrower had vs what was assigned to him. A high difference may indicate taking up more loans in the future.
QuotDefaults	Decimal numbers	numeric	<p>Total number of credits until date x. A Quotient is calculated for each loan of a borrower.</p> <p>Example: Borrower is having 3 loans. 1 defaulted, the quotient at the time of 3 total loans for defaulting is 0.33. When a 4<sup>th</sup> loan is taken and no further default has taken place, the quotient is lowered to 0.25</p> <p>Finally be advised that the number of loans is not taken from the value in variable “NoOfPreviousLoansBeforeLoan” but from the number of loans per borrower in the LoanData.csv-file.</p>

Independent Feature Variablename	Possible values	Vartype	Calculation; Motivation
DaysLastCred	Whole numbers	numeric	LoanDate; Difference of time in days between a borrowers credits. First credit is always zero, difference for second, third, e.g. is calculated by difference of consecutive dates
MMYY_credit	Format - "%m-%Y"	date	Calculated from "LoanDate" - recoded to month_year  Used for graphical presentations to ensure that time series graphs are possible – a mere helper-variable

### 3.3 Creation of Loan Analysis Dataset – Filtering out further insignificant data

From the original LoanData data-set a Loan Analysis data-set is created. This will be the data we are running our logistic regression on. The most important filter in defining what is part of the analysis set is the the following criteria:

→ "MaturityDate\_Last" < "ReportAsOfEOD" (The loan has to be finished!)

Concerning the filter "MaturityDate\_Last" < "ReportAsOfEOD" it is imperative that the loan has ended (matured) and no more payment dates are available for said loan. The loan has come to full fruition and we are interested in the value of the variable "WorseLateCategory" - if it shows a categorical value higher than 60 days we had defined the dependent variable "Defaulted" as "1" - or TRUE if you will. Would there still be any payment periods open for a loan, the "WorseLateCategory" may still change to defaulted. That is why only finalized loans are taken into consideration for the analysis data-set! Of course this lowers the count of available cases to a low count (see following table) but it assures the data correctness as a change of "WorseLateCategory" should not be possible anymore. Be advised that this has a very significant effect on the available number of analysis data as the older a credit is, the more likely it has already been finished. 30911 out of 150623 rows remain after activating this filter. And they had to be cut down even more as one will see in the following paragraphs!

### 3.4 Data cleaning, presumed Errors and outlier detection

Further Investigation reveals some problems with the analysis data-set, which hold true for the LoanData-set also.

These problematic variables are:

- DebtToIncome
- FreeCash
- RefinanceLiabilities
- PreviousEarlyRepaymentsBefoleLoan
- Amount vs. ManualBids
- MonthlyPayment

How I handled these presumed errors is shown in the following paragraphs.

#### 3.4.1 Missing data for multiple variables after 2017

To identify possible outliers or strange data behavior the independent variables have not only been plotted as a simple boxplot, but also vs the calculated feature variable "MMYY\_credit". A full overview of all graphs can be found in the Appendix – here only notable findings are being discussed.

For the following variables a certain discontinuation of data seems to have taken place. We start with the analysis of "DebtToIncome". The area in question is marked by a red box with dashed lines. It is rather improbable that the "DebtToIncome" suddenly drops to zero. The cause for the sudden drop is that most of the values are missing starting from July 2017.

This is also true for the variables:

- DebtToIncome
- FreeCash

Be aware: This document and the information contained within does not substitute financial advice from professionals.

### → RefinanceLiabilities

As it can only be guessed, why the values are missing (discontinuation of data collection, error in db, error in data extraction, e.g.), I decided to not use these variables at all and to drop them from the list of independent variables! An alternative would be the imputation of values either by extrapolating missing values by regression or averaging existing values but this possibility has been disregarded because not singular months have been lost, but a whole time period like a full year which makes approximative extrapolation rather shaky.

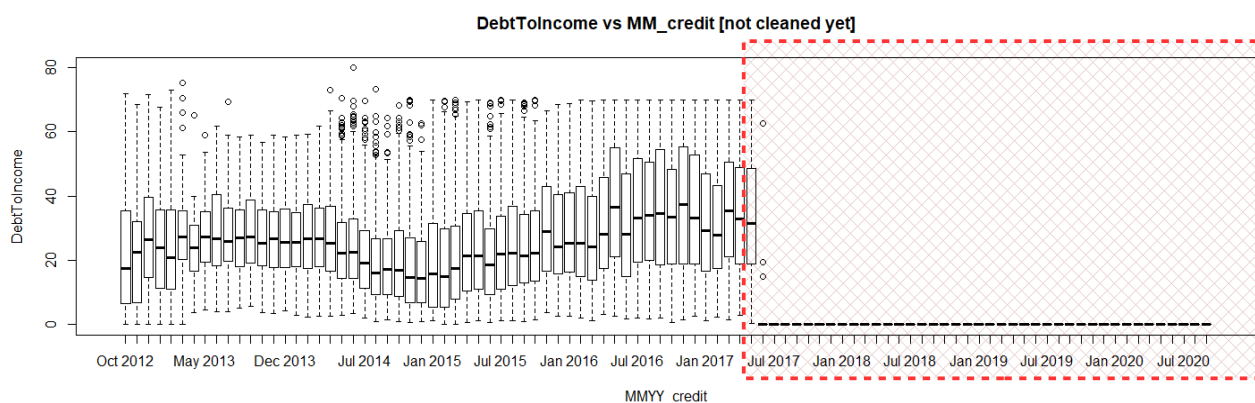


Figure 3.2: Missing values for DebtToIncome - See red marked box to the right

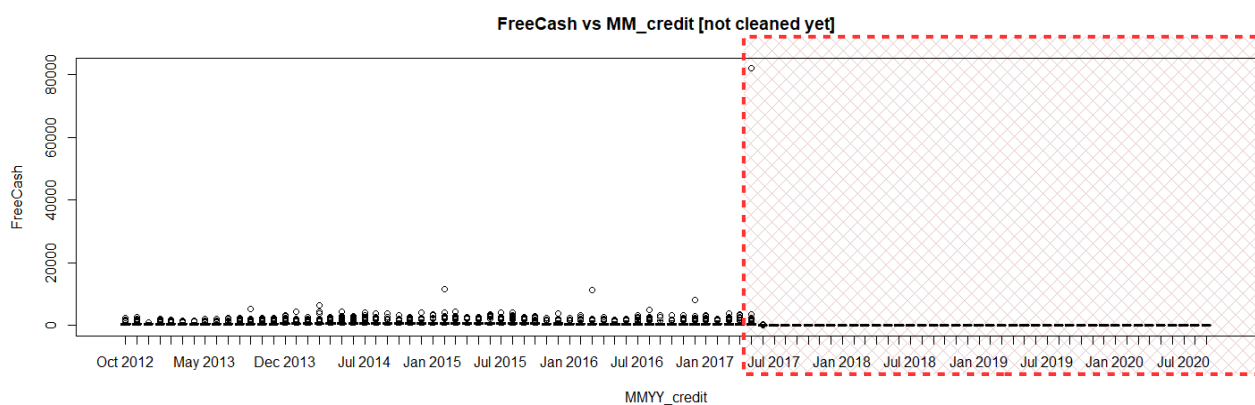


Figure 3.3: Missing values for FreeCash - See red marked box to the right

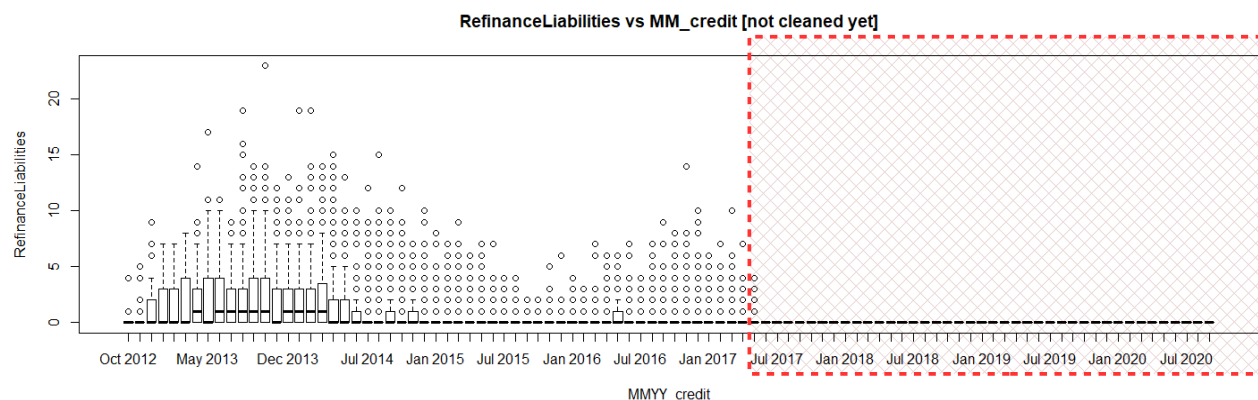


Figure 3.4: Missing values for RefinanceLiabilities - See red marked box to the right

For the variable "PreviousEarlyRepaymentsBefoleLoan" there seems to be a steady decline in available data after May 2018. Before that it is either Missing or zero. The following table (a boxplot graph is not showing the extent of missing values) is a simple overview of missing values (TRUE) per month. As can be seen, until the April of 2018, no missing values are given – starting May 2018 there is a stark change in number of missing values - they are drastically increasing. As I cannot make out a good reason for this, the whole variable will be removed from the collection of independent variables.

	Feb 2009	Mar 2009	Apr 2009	May 2009	Jun 2009	Jul 2009	Aug 2009	Sep 2009	Oct 2009	Nov 2009	Dec 2009	Jan 2010	Feb 2010	Mar 2010	Apr 2010	May 2010	Jun 2010	Jul 2010	Aug 2010	Sep 2010	Oct 2010
FALSE	1	37	41	55	50	67	89	79	82	69	95	76	126	192	132	106	86	76	72	82	79
TRUE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Nov 2010	Dec 2010	Jan 2011	Feb 2011	Mar 2011	Apr 2011	May 2011	Jun 2011	Jul 2011	Aug 2011	Sep 2011	Oct 2011	Nov 2011	Dec 2011	Jan 2012	Feb 2012	Mar 2012	Apr 2012	May 2012	Jun 2012	Jul 2012
FALSE	72	58	40	32	54	48	46	50	29	38	33	27	31	23	40	26	20	31	26	18	31
TRUE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Aug 2012	Sep 2012	Oct 2012	Nov 2012	Dec 2012	Jan 2013	Feb 2013	Mar 2013	Apr 2013	May 2013	Jun 2013	Jul 2013	Aug 2013	Sep 2013	Oct 2013	Nov 2013	Dec 2013	Jan 2014	Feb 2014	Mar 2014	Apr 2014
FALSE	22	20	58	74	88	94	99	110	119	133	143	224	254	262	323	364	385	528	402	437	468
TRUE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	May 2014	Jun 2014	Jul 2014	Aug 2014	Sep 2014	Oct 2014	Nov 2014	Dec 2014	Jan 2015	Feb 2015	Mar 2015	Apr 2015	May 2015	Jun 2015	Jul 2015	Aug 2015	Sep 2015	Oct 2015	Nov 2015	Dec 2015	Jan 2016
FALSE	525	601	652	784	762	830	785	681	906	788	713	648	528	537	729	683	625	576	683	630	602
TRUE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Feb 2016	Mar 2016	Apr 2016	May 2016	Jun 2016	Jul 2016	Aug 2016	Sep 2016	Oct 2016	Nov 2016	Dec 2016	Jan 2017	Feb 2017	Mar 2017	Apr 2017	May 2017	Jun 2017	Jul 2017	Aug 2017	Sep 2017	Oct 2017
FALSE	725	851	822	818	832	920	1092	771	749	1155	1177	1137	845	1112	1055	1164	1347	1723	2063	1978	1837
TRUE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Nov 2017	Dec 2017	Jan 2018	Feb 2018	Mar 2018	Apr 2018	May 2018	Jun 2018	Jul 2018	Aug 2018	Sep 2018	Oct 2018	Nov 2018	Dec 2018	Jan 2019	Feb 2019	Mar 2019	Apr 2019	May 2019	Jun 2019	Jul 2019
FALSE	1873	1799	1888	1542	1594	1697	675	53	33	61	35	67	72	51	91	62	58	70	69	81	75
TRUE	0	0	0	0	0	0	1056	1790	1971	1937	1930	2928	2708	3015	3026	2594	3259	3575	3640	3397	5706
	Aug 2019	Sep 2019	Oct 2019	Nov 2019	Dec 2019	Jan 2020	Feb 2020	Mar 2020	Apr 2020	May 2020	Jun 2020	Jul 2020	Aug 2020	Sep 2020	Oct 2020						
FALSE	87	80	117	117	80	108	71	67	35	28	14	33	28	3							
TRUE	5295	5613	7244	5870	5630	5252	4630	4119	1414	1264	906	1199	1162	166							

Figure 3.5: Start of Missing Values "NA" for "PreviousEarlyRepaymentsBefoleLoan" - starting May 2018 and ever increasing

### 3.4.2 Treatment of outliers

Concerning outliers there seem to be a significant ones all over the used independent variables. For example "Free Cash" provides one value with well over 80000€. This may be an error and seems rather outlandish but I decided to be not to quick in deleting any values even when they lie out of the  $\pm 1.5 \times Q1/Q3$  interquartile-range bounds (above/below the "whiskers") as there is no way to surely say that these are errors or real values. So they will be kept.

### 3.4.3 Strange Correlations between Amount and ManualBids for data before October 2012

A special case of seemingly erroneous data seems to be the near perfect correlation between the variables of "Bids-Manual" and "Amount" for a certain time-frame. I came across this when eyeballing the data and found it strange that the value of "BidsManual" were ever so often very close or even identical to the value of "Amount". Maybe it was due to the nature of both variables – so I consulted the definitions given on the Bondora web-page<sup>1</sup>.

According to the variable definition given on the Bondora web-page "BidsManual" is documenting the:

*"The amount of investment offers made manually"*

<sup>1</sup> <https://www.bondora.com/en/public-reports> (09/09/2020)

which implicates that it counts the manual offers made via Bondora Web-app to obtain investment in the credit.

The variable "Amount" does define the:

*"Amount the borrower received on the Primary Market. This is the principal balance of your purchase from Secondary Market "*

If that holds true, I do not see any relationship between "BidsManual" and "Amount" - therefore the closeness of each value is kind of odd. To further investigate I did an analysis of mean-correlation coefficients with "MMYY\_credit" as the grouping variable to see if there is any correlation (correlation does not mean causal relationship – agreed, but it might point out if values move in unison).

The results are shown in the following graph – see the red box with the dashed line for high correlation coef.:

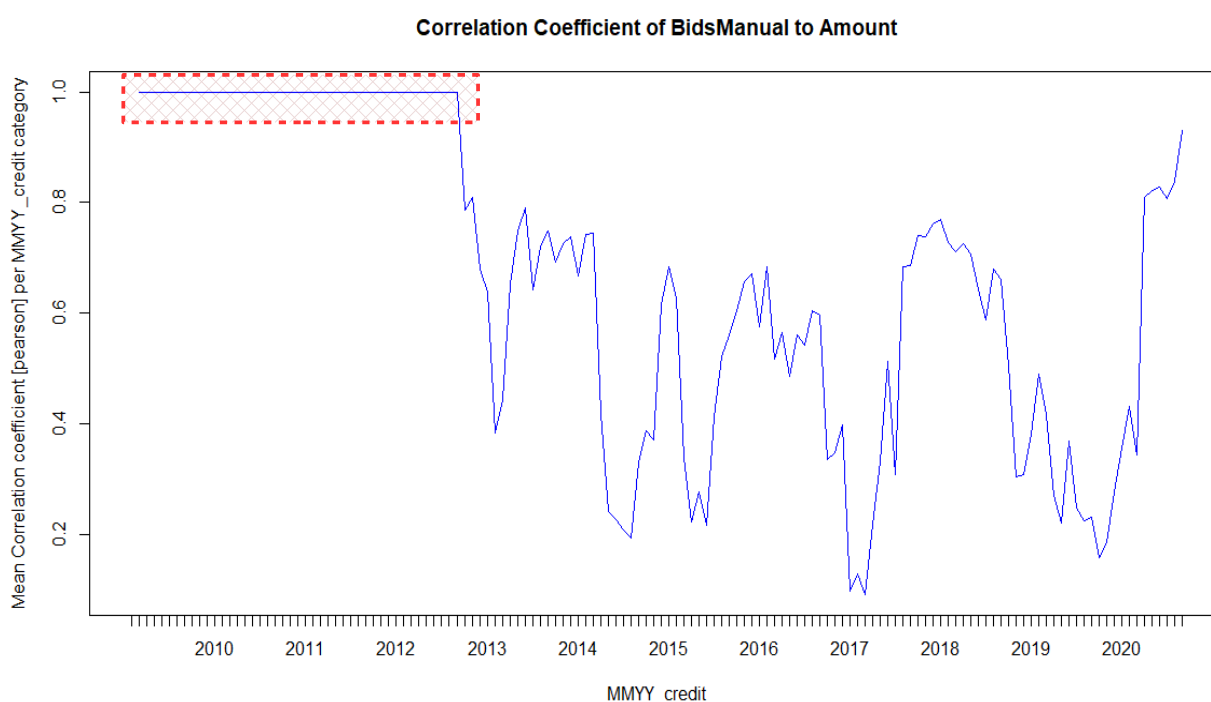


Figure 3.6: There is an unusual high correlation of "BidsManual" to "Amount" until mid 2012

The red marked zone shows the perfect correlation value of one (in the mean). This must not necessarily be an indicator for erroneous values as the values of both variables may move in the same direction but may show a big absolute difference, so I decided to check the absolute mean of the values per "MMYY\_credits" per variable. Until Oct 2012 the differences between "BidsManual" and "Amount" are minuscule. Either there is an error in the data Transfer of the Loan Data csv-file (for example values of BidsManual have been also copied to Amount) or there is a misinterpretation by me.

A small selection of the correlation and mean values of respective variables is shown here:

MMYY_credit	Correlation_BidsManual_Amount	Mean_BidsManual	Mean_Amount
Feb 2009	NA	322.753900	322.7544
Mar 2009	1.00000000	123.936765	123.9360
Apr 2009	1.00000000	107.870378	107.8700
May 2009	1.00000000	161.289738	161.2894
Jun 2009	0.99999999	188.539336	188.5387

MMYY_credit	Correlation_BidsManual_Amount	Mean_BidsManual	Mean_Amount
Jul 2009	0.99999998	220.924425	220.9252
Aug 2009	0.99999999	232.020761	232.0209
Sep 2009	0.99999998	224.580652	224.5773
Oct 2009	0.99999998	229.926033	229.9260
Nov 2009	0.99999999	198.774467	198.7741
...	...	...	...

As I cannot decide if this is an data-error or misinterpretation by me I am willing to discard all data in the Loan analysis data-set until August 2012! This enables me to continue to use the variable "Amount" in the further analysis but I lose the years 2009 till August 2012 for all variables in the Loan analysis completely.

### 3.4.4 Monthly Payment always zero or missing until mid 2014

Also for the variable "MonthlyPayment" notable differences for the start of the time-period (2009 until 2014) have been found. The first value other than zero or "NA" can be first found for the Middle of 2014, only then numbers higher then zero are detected. See the following graphic.

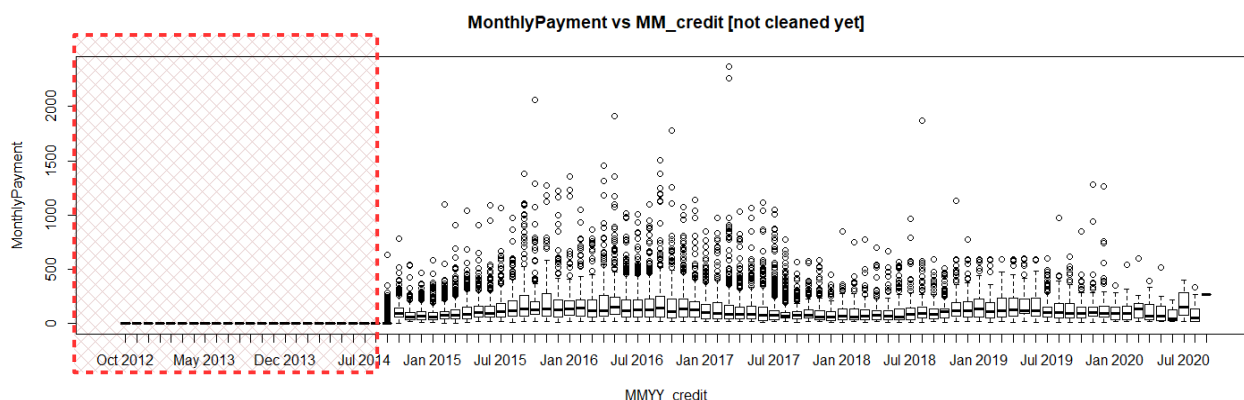


Figure 3.7: Missing values for MonthlyPayment - See red marked box to the left

This did not make sense to as me, there should always be monthly payments by the borrower, so I decided to only take rows into the analysis, which show values higher then zero or are unlike "NA" for "MonthlyPayment":

→ "MonthlyPayment" > 0

The deletion of seemingly erroneous data started out with 26396 rows. But by adding multiple filters as described above I finally ended up with 19450 rows, which means 6946 rows deleted in the analysis data because of missing data or seemingly erroneous ones.

### 3.4.5 Deletion of low count Levels of Factor

Concerning the deletion of all factor levels which sum of appearance is lower than 50 → this was done as I tried to avoid lowly covered factor levels which may infer with glm-regression itself or overweight a certain factor by possible low variety. As an example the R-code created a table of number of appearances of the respective level codes of the factorial variable and delete any level which sum is small than 50. In the Loan analysis data for country there are 4 levels: EE; ES ; FI; SK. For none of them the count was lower than 50 therefore they remained in the analysis.

## 4. Building the Logistic Model

### 4.1 Test of significance

After identifying probable errors and avoiding them by either not using the variable itself or cutting time-ranges from the analysis altogether a test of significance has been done. As it stands, the following independent variables were still in the model before the test of significance (removed variables due to data errors have been crossed out):

Independent Variablename	Possible values	Vartype
NewCreditCustomer	TRUE/FALSE	logical
LanguageCode	Multiple levels	factor
Age	Whole numbers	numeric
Gender	Multiple levels	factor
Country	Multiple levels	factor
AppliedAmount	Decimal numbers	numeric
FreeCash	Decimal numbers	numeric
DebtToIncome	Decimal numbers	numeric
Amount	Decimal numbers	numeric
Interest	Decimal numbers	numeric
LoanDuration	Whole numbers	numeric
MonthlyPayment	Decimal numbers	numeric
Education	Multiple levels	factor
EmploymentDurationCurrentEmployer	Multiple levels	factor
HomeOwnershipType	Multiple levels	factor
IncomeTotal	Decimal numbers	numeric
ExistingLiabilities	Whole numbers	numeric
LiabilitiesTotal	Decimal numbers	numeric
RefinanceLiabilities	Whole numbers	numeric
NoOfPreviousLoansBeforeLoan	Whole numbers	numeric
AmountOfPreviousLoansBeforeLoan	Decimal numbers	numeric
PreviousRepaymentsBeforeLoan	Decimal numbers	numeric
PreviousEarlyRepaymentsBeforeLoan	Decimal numbers	numeric
PreviousEarlyRepaymentsCountBeforeLoan	Whole numbers	numeric

The remaining variables are tested for collinearity and predictory significance.

## 4.2 Detection of collinearity – and its avoidance

Before a simple logistic regression has been run a correlation table is inspected to identify possible correlations between the independent variables. As can be seen from the following graph, there are some of these correlations to be found (we define  $r > 0.5$  as relevant) – see red box with dashed line:

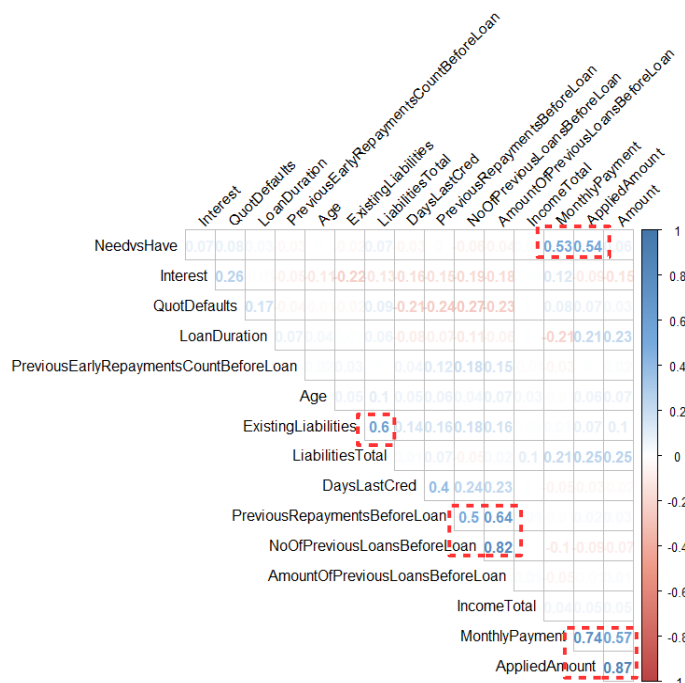


Figure 4.1: Correlations  $> +/- 0.5$  highlighted by red box

We remove the following variables to avoid collinearity and an artificial inflation of variance:

- ➔ remove "AppliedAmount" – correlated with "Amount" – I do believe Amount is more relevant so I am keeping it
- ➔ remove "NoOfPreviousLoansBeforeLoan" – correlated with "AmountOfPreviousLoansBeforeLoan" – I like to keep the variable with more detailed information. An info on the total amount of the previous loans is more graduated than the simple number on the total number of loans itself.
- ➔ remove "PreviousRepaymentsBeforeLoan" – correlated with "NoOfPreviousLoansBeforeLoan" and "AmountOfPreviousLoansBeforeLoan"
- ➔ remove "ExistingLiabilities" – correlated with "LiabilitiesTotal" – I like to keep the variable with more detailed information. An info on the total amount of the "LiabilitiesTotal" (range 0-12400000) is more graduated than the simple number on the total number of loans itself, as given with "ExistingLiabilities" (range 0-40).
- ➔ remove "MonthlyPayment" – correlated with "Amount" and "AppliedAmount" – though one – both are good indicators of the financial pressure residing on the borrower. I decide to remove "MonthlyPayment" from the list of independent variables as this might change by changing the duration or interest of the loan. "Amount" seems to be the more stable variable.



The correlation matrix after removal of highly correlated variables:

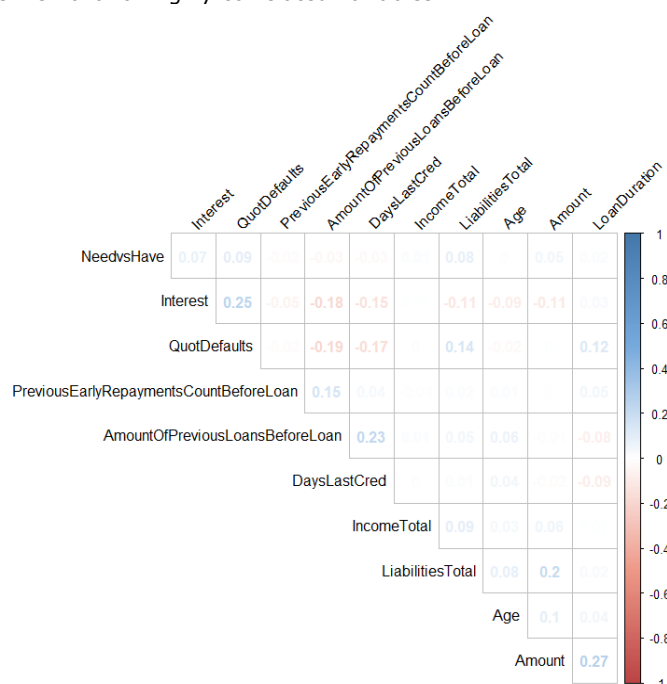


Figure 4.2: Correlation matrix after removing strongly correlated variables

#### 4.3 Running a simple glm – how significant are the independent variables

After cleaning out the list of independent variables by making sure no erroneous, or highly correlated variables are used anymore, a simple glm-regression has been run to obtain the significance each variable provides in explaining the variance. If an independent variable is shown to be insignificant – it will be removed from the model as it does not provide to the explanation of variance of the model and will be taken out for sake of parsimony. The dependent variable is, as known and remains, "Defaulted".

As can be seen on the following pages a part of the `r summary()`-function has been obtained which informs about the significance each variable is providing in explaining the variance of the model.

Be advised that I did not remove "Interest" from the model, even if it does not show up as a significant variable with a p-value of 0.2261. I believe it is a very important variable in estimating the probability of default. The low p-value seems to be an artifact of the selected independent variables (which is true as later will be discovered after removing other variables).

The following variables have been found to have no significant explanatory contribution and are therefore removed from the model:

- ➔ removed "Country"
- ➔ removed "Age"
- ➔ removed "Gender"
- ➔ removed "EmploymentDurationCurrentEmployer"
- ➔ removed "NeedsvsHave"
- ➔ removed "LanguageCode"
- ➔ removed "Education"
- ➔ removed "PreviousEarlyRepaymentsBeforeLoan"
- ➔ removed "AmountOfPreviousLoansBeforeLoan"
- ➔ removed "PreviousEarlyRepaymentsCountBeforeLoan"
- ➔ removed "HomeOwnershipType"
- ➔ removed "LiabilitiesTotal"




## Coefficients:





	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-5.969e+00	7.001e-01	-8.525	< 2e-16	***
NewCreditCustomerTRUE	-1.127e+00	2.496e-01	-4.516	6.30e-06	***
LanguageCode2	1.164e+00	9.084e-01	1.281	0.2001	
LanguageCode3	3.597e-01	3.199e-01	1.124	0.2610	
LanguageCode4	2.402e-01	1.133e+00	0.212	0.8321	
LanguageCode6	3.834e-02	2.293e+00	0.017	0.9867	
LanguageCode9	-6.957e+00	5.354e+02	-0.013	0.9896	
Age	-2.672e-03	8.848e-03	-0.302	0.7626	
Gender1	-1.623e-01	1.885e-01	-0.861	0.3893	
CountryES	9.981e-01	2.299e+00	0.434	0.6641	
CountryFI	-6.783e-01	1.124e+00	-0.604	0.5461	
CountrySK	8.892e+00	5.354e+02	0.017	0.9867	
Amount	1.112e-04	4.488e-05	2.476	0.0133	*
Interest	5.612e-03	4.636e-03	1.211	0.2261	
LoanDuration	2.092e-02	5.210e-03	4.015	5.95e-05	***
Education2	2.235e-02	5.544e-01	0.040	0.9678	
Education3	-3.046e-01	5.143e-01	-0.592	0.5537	
Education4	-1.610e-01	4.922e-01	-0.327	0.7437	
Education5	-5.918e-01	4.989e-01	-1.186	0.2355	
EmploymentDurationCurrentEmployerOther	-1.221e+00	1.334e+00	-0.915	0.3602	
EmploymentDurationCurrentEmployerRetiree	-1.824e+00	9.703e-01	-1.880	0.0601	.
EmploymentDurationCurrentEmployerTrialPeriod	-1.858e+00	1.088e+00	-1.708	0.0877	.
EmploymentDurationCurrentEmployerUpTo1Year	2.109e-01	2.529e-01	0.834	0.4042	
EmploymentDurationCurrentEmployerUpTo2Years	1.458e-01	2.956e-01	0.493	0.6218	
EmploymentDurationCurrentEmployerUpTo3Years	1.341e-01	3.226e-01	0.416	0.6776	
EmploymentDurationCurrentEmployerUpTo4Years	3.622e-01	3.442e-01	1.052	0.2926	
EmploymentDurationCurrentEmployerUpTo5Years	-4.529e-01	2.751e-01	-1.646	0.0997	.
HomeOwnershipType2	-5.304e-01	2.750e-01	-1.929	0.0538	.
HomeOwnershipType3	-1.083e-01	2.474e-01	-0.438	0.6615	
HomeOwnershipType4	-3.499e-02	3.779e-01	-0.093	0.9262	
HomeOwnershipType5	3.992e-01	6.318e-01	0.632	0.5275	
HomeOwnershipType6	2.323e-01	5.159e-01	0.450	0.6525	
HomeOwnershipType7	2.224e-01	4.316e-01	0.515	0.6063	
HomeOwnershipType8	6.998e-02	3.102e-01	0.226	0.8215	
HomeOwnershipType9	-7.745e-01	8.863e-01	-0.874	0.3822	
HomeOwnershipType10	-1.583e+00	8.863e-01	-1.786	0.0741	.
IncomeTotal	-2.992e-05	1.380e-05	-2.168	0.0302	*
LiabilitiesTotal	-1.280e-04	1.431e-04	-0.895	0.3709	
AmountOfPreviousLoansBeforeLoan	-3.988e-05	2.427e-05	-1.643	0.1004	
PreviousEarlyRepaymentsCountBeforeLoan	1.488e-02	1.233e-01	0.121	0.9040	
NeedysHave	5.482e-05	8.817e-05	0.622	0.5342	
QuotDefaults	1.378e+01	3.888e-01	35.443	< 2e-16	***
DaysLastCred	-8.792e-04	2.073e-04	-4.241	2.23e-05	***
---					

signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Figure 4.3: r-summary() report of full statistical model

Interpretation of the remaining variables regression signs. The regression coefficient values itself will not be reported.

Independent significant Variable-name	Sign-Interpretation	Risk of Defaulting
NewCreditCustomer	Negative – Therefore the probability of a loan defaulting decreases for a new credit customer	
Amount	Positive – Therefore for each € of loan amount, the probability of a loan defaulting increases	
Interest	Positive – Therefore for each percentage point, the probability of a loan defaulting increases	

Independent significant Variable-name	Sign-Interpretation	Risk of Defaulting
LoanDuration	Positive – Therefore for each month of duration of a loan, the probability of a loan defaulting increases	
IncomeTotal	Negative – Therefore for each € of IncomeTotal of a borrower the probability of a loan defaulting decreases	
QuotDefaults	Positive – Therefore if the quotient is not zero (aka the borrower had at least one defaulted loan) the probability of a defaulting loan is increasing.	
DaysLastCred	Negative – Therefore the longer the last credit lies behind the borrower (in days), the more the probability of defaulting is lowered.	

#### 4.4 Calculating McFaddens Pseudo-R<sup>2</sup> – Stepwise addition of independent variables and testing prediction capability

After deleting insignificant variables from the model an analysis of the explained variance is made which might tell us about the ability of our model to provide good enough predictions outside the used training data-set. As we are using a logistical regression McFaddens Pseudo-R<sup>2</sup> is used to obtain a statistical measure which can be interpreted likewise to the R<sup>2</sup> in lets say linear regressions. By looping over the pool of still available independent variables we are adding them one by one to the formula, each time calculating McFaddens Pseudo Pseudo-R<sup>2</sup>.

The Loan Analysis data-set is divided into 2 different parts – the **trainer data-set**, consisting of 70 % of all available Analyses data-sets rows and a **test data-set**, consisting of the other 30 %. The trainer data-set will not only be used to calculate McFaddens Pseudo-R<sup>2</sup> but I also will use the regression coefficients derived from the training data to predict the probability of default in the test data-set. These predictions will be compared to the true historical defaulting status of the loan in the test data-set . By this the application of the statistical model and its usage in correctly predicting defaulting risk can be tracked.

The most important results an be found in the following table:

Model-#	Added variable (Added to model run before!)	McFaddens Pseudo-R <sup>2</sup>	Errors in prediction of test data-set	Hint/Note
1	NewCreditCustomer	'log Lik.' 0.07321966 (df=2)	39,9 % wrong	<b>Low Pseudo-R<sup>2</sup></b>
2	Model 1 + Amount	'log Lik.' 0.07337988 (df=3)	39,9 % wrong	<b>Low Pseudo-R<sup>2</sup></b>
3	Model 2 + Interest	'log Lik.' 0.1034949 (df=4)	37,03 % wrong	<b>Low Pseudo-R<sup>2</sup></b>
4	Model 3 + LoanDuration	'log Lik.' 0.1113993 (df=5)	38,6 %wrong	<b>Low Pseudo-R<sup>2</sup></b>
5	Model 4 + IncomeTotal	'log Lik.' 0.1113998 (df=6)	38,6 % wrong	<b>Low Pseudo-R<sup>2</sup></b>
6	Model 5 + QuotDefaults	'log Lik.' 0.9554205 (df=7)	1,4 % wrong	<b>Overfitting!</b>
7	Model 6 + DaysLastCred	'log Lik.' 0.9866278 (df=8)	1,3 % wrong	<b>Overfitting!</b>

As can be seen from the table the predictory power of the model concerning the test data-set is very low until "QuotDefaults" is added. Yet this addition leads to such an inflation of predictory power it can only be assumed as "overfit-

ted". The sudden and significant rise in variance explained seems like a warning signal to me, rather than a golden bullet of the perfect statistical model to predict defaulting credits.

```

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.402e+00  3.465e-01 -18.474 < 2e-16 ***
NewCreditCustomerTRUE -1.298e+00  2.667e-01  -4.866 1.14e-06 ***
Amount      8.918e-05  5.059e-05   1.763 0.077926 .
Interest    1.305e-02  3.710e-03   3.517 0.000437 ***
LoanDuration 1.567e-02  6.036e-03   2.597 0.009408 **
IncomeTotal -4.240e-05  2.065e-05  -2.053 0.040065 *
QuotDefaults 1.374e+01  4.440e-01  30.948 < 2e-16 ***
DaysLastCred -1.038e-03  2.275e-04  -4.563 5.04e-06 ***
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 18398.46  on 13610  degrees of freedom
Residual deviance:   838.91  on 13603  degrees of freedom
AIC: 854.91

```

```
Number of Fisher Scoring iterations: 9
```

Figure 4.4: `r-summary()` model #7

A graphical representation of the relative Importance of independent variables in explaining the defaulting probability in the statistical model used:

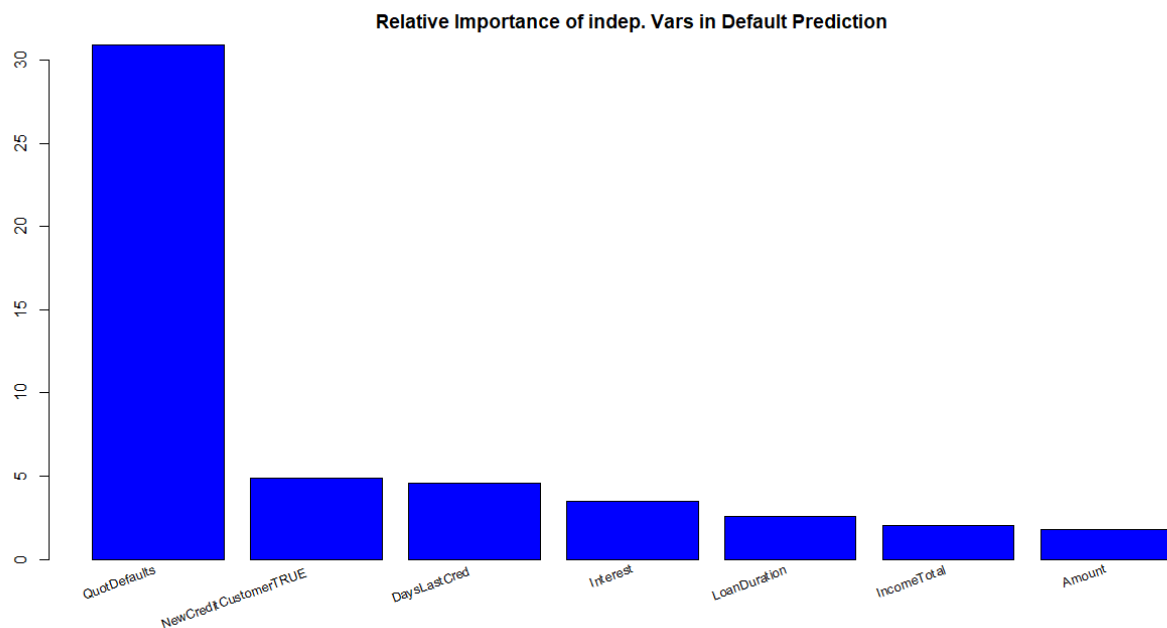


Figure 4.5: Relative Importance of independent Variables in model #7

#### 4.5 Graphical depiction of predictions with model #5 and #7

As pointed out I predicted the possibility of defaulting loans using each statistical model on the Loan Analysis test data-set. For statistical model #5 and #7 the predictions results, named "PredictionProbabilityOfDefault", have been plotted against the Bondora provided variable "ProbabilityOfDefault". I did this to see, how far my calculated values are away from Bondora's given defaulting probability. Do not misunderstand this as criticism or endorsement of Bondora's "ProbabilityOfDefault" variable but as a real world test, something to compare to – there is no judgment of the Bondora provided variable intended. As will be seen, model #5 tends to be more indecisive than Bondora's probability estimation, while model #7 does not know any shades of grey and occupies either default or non-defaulting positions.

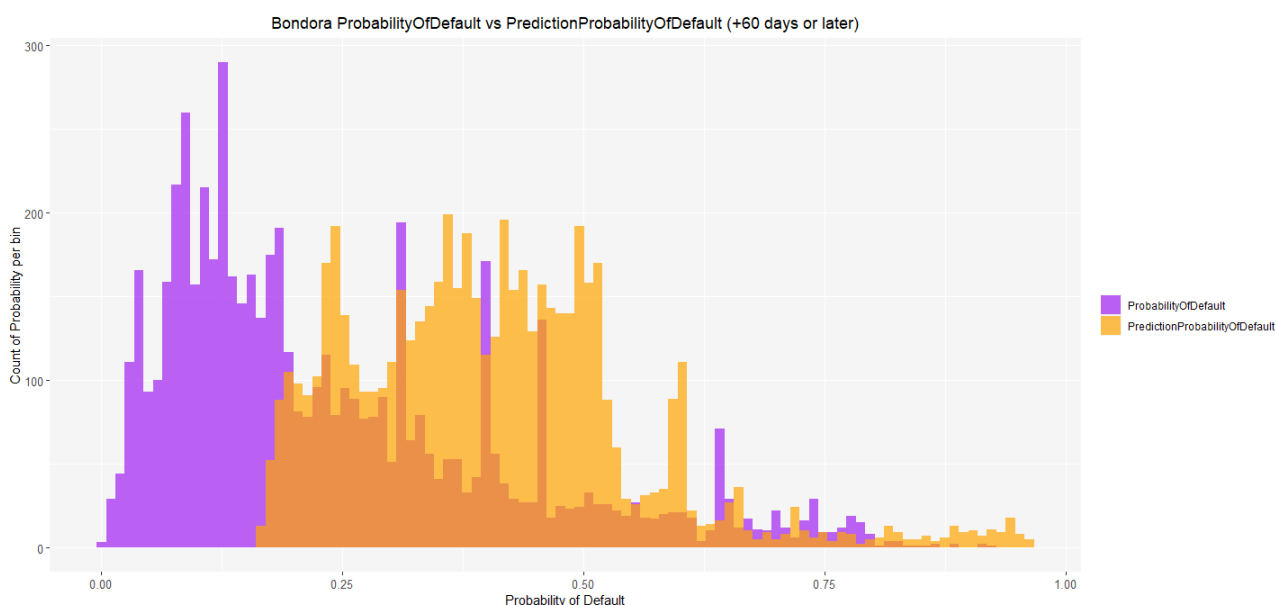


Figure 4.6: Using Model #5 for prediction of probability of defaulting – see orange values for predicted, purple is Bondora given probability

As can be seen from the above graph statistical model #5 shows a more natural distribution of probability of defaulting than model 7. It tends to be less decisive than Bondora's defaulting estimate, as it seems to center with most of its values between probabilities of 0.15 and 0.65.

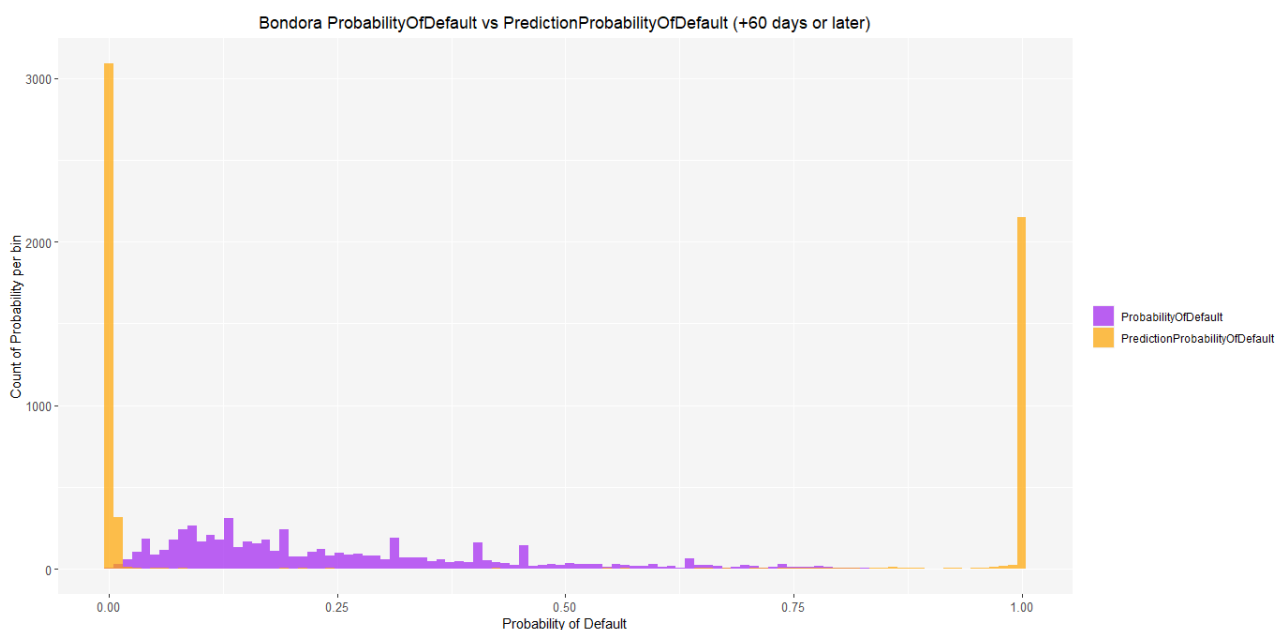


Figure 4.7: Using Model #7 for prediction of probability of defaulting – see orange values for predicted, purple is Bondora given probability

Be aware: This document and the information contained within does not substitute financial advice from professionals.

Be advised that in figure 4.7 identical data is used as in figure 4.6. The values of Bondora's Probability of Default are just visually compressed as the values of the predicted Defaulting probability are occupying both extreme ends. From the above graph it becomes clear that model #7 does not provide any fine tuned differentiation on defaulting probability prediction. It is either located at the far "0" - aka non-defaulting, or "1" - aka defaulting side. This seems unnatural and to be an effect of overfitting the statistical model.

## 5. Final note on the results

---

It seems that from the data provided by the Loan Data-set a model which is able to predict a default or non-defaulting of a loan securely is barely possible. In the course of this document after cleaning and analyzing the data two models have been created which could be used to predict this probability.

Model #5 using basically unaltered original variables from the data-set and model #7, which contains 2 Feature calculated variables plus the existing, original variables in model #5. While model #7 does show a high percentage of explained variance it just seems to be too much – too good to be true. It surely is overfitted as it only gives back 0 or 1 (with few exceptions). Such a “black and white” result without shades of gray is unrealistic. Model #5 is “more tempered” in its results and shows an even distribution of defaulting probabilities yet the low value of explained Pseudo- $R^2$  (11.13 %) makes the reliability of the predictions rather low.

The addition of variables in the data-set, that are already providing a risk evaluation of defaulting probability would most certainly have helped but as pointed out – the target of this analysis was to do without external risk assessment and to come up with an own model of risk evaluation.

It also may be that the variable(s) needed for a better explanation of the variance are not to be found in Bondora's Loan Data-set. External data like industrial indices, number of unemployed in Eastern Europe or stock-market indices could improve the situation.

Finally - a lot of analysis data got lost because of assumed errors in the raw-data. Maybe other results are found, if the remaining Sample Size after data cleaning could be increased.

## 6. Appendix

### 6.1 Raw Data – Data Types after initial import

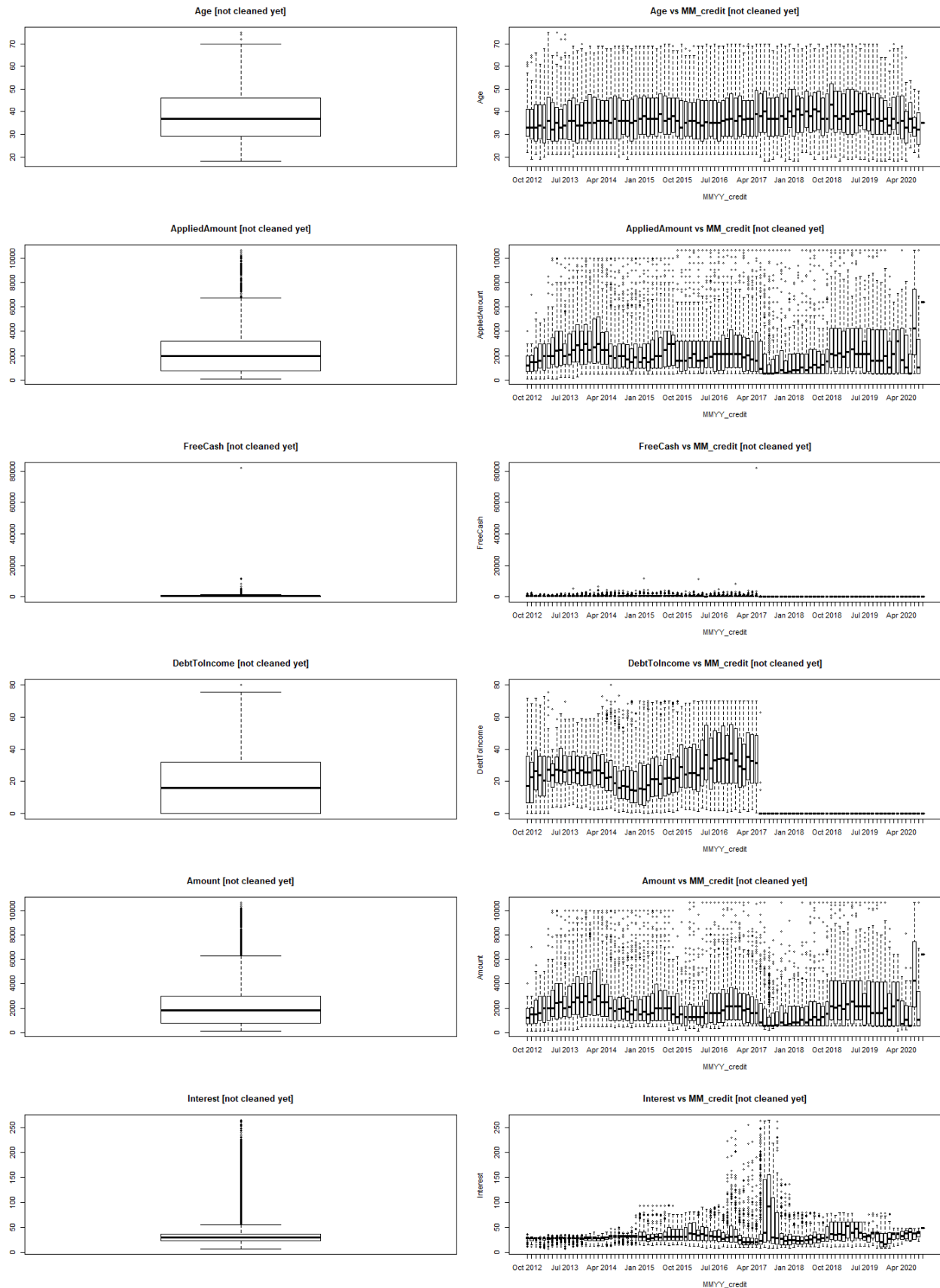
This table informs about the initial data type before they were reformatted.

Variable Name	Variable Type	Variable Name	Variable Type	Variable Name	Variable Type
ReportAsOfEOD	Date	Education	numeric	DebtOccuredOn-ForSecondary	Date
LoanId	character	MaritalStatus	numeric	ExpectedLoss	numeric
LoanNumber	numeric	NrOfDependants	numeric	LossGivenDefault	numeric
ListedOnUTC1	POSIXct	EmploymentStatus	numeric	ExpectedReturn	numeric
ListedOnUTC2	POSIXt	EmploymentDurationCurrentEmployer	character	ProbabilityOfDefault	numeric
BiddingStartedOn1	POSIXct	EmploymentPosition	logical	DefaultDate	Date
BiddingStartedOn2	POSIXt	WorkExperience	character	PrincipalOverdue-BySchedule	numeric
BidsPortfolioManager	numeric	OccupationArea	numeric	PlannedPrincipal-PostDefault	numeric
BidsApi	numeric	HomeOwnershipType	numeric	PlannedInterestPost-Default	numeric
BidsManual	numeric	IncomeFromPrincipalEmployer	numeric	EAD1	numeric
UserName	character	IncomeFromPension	numeric	EAD2	numeric
NewCreditCustomer	logical	IncomeFromFamilyAllowance	numeric	PrincipalRecovery	numeric
LoanApplication-StartedDate1	POSIXct	IncomeFromSocial-Welfare	numeric	InterestRecovery	numeric
LoanApplication-StartedDate2	POSIXt	IncomeFromLeave-Pay	numeric	RecoveryStage	numeric
LoanDate	Date	IncomeFromChild-Support	numeric	StageActiveSince1	POSIXct
ContractEndDate	Date	IncomeOther	numeric	StageActiveSince2	POSIXt
FirstPaymentDate	Date	IncomeTotal	numeric	ModelVersion	numeric
MaturityDate_Original	Date	ExistingLiabilities	numeric	Rating	character
MaturityDate_Last	Date	LiabilitiesTotal	numeric	EL_V0	logical
ApplicationSigned-Hour	numeric	RefinanceLiabilities	numeric	Rating_V0	logical
ApplicationSigned-Weekday	numeric	DebtToIncome	numeric	EL_V1	numeric
VerificationType	numeric	FreeCash	numeric	Rating_V1	character
LanguageCode	numeric	MonthlyPaymentDay	numeric	Rating_V2	character
Age	numeric	ActiveScheduleFirst-PaymentReached	logical	Status	character
DateOfBirth	logical	PlannedPrincipalTill-Date	numeric	Restructured	logical
Gender	numeric	PlannedInterestTill-Date	numeric	ActiveLateCategory	character
Country	character	LastPaymentOn	Date	WorseLateCategory	character
AppliedAmount	numeric	CurrentDebtDaysPrimary	numeric	CreditScoreEsMicroL	character
Amount	numeric	DebtOccuredOn	Date	CreditScoreEsE-quirifaxRisk	character
Interest	numeric	CurrentDebt-DaysSecondary	numeric	CreditScoreFiAsi-akasTietoRiskGrade	character
LoanDuration	numeric			CreditScoreEeMini	numeric
MonthlyPayment	numeric			PrincipalPayments-Made	numeric
County	logical			InterestAndPenalty-PaymentsMade	numeric
City	logical				
UseOfLoan	numeric				

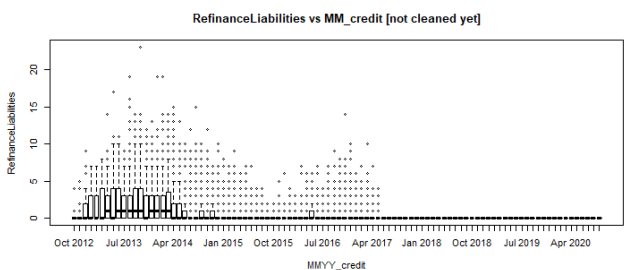
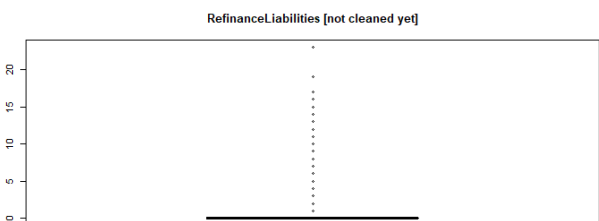
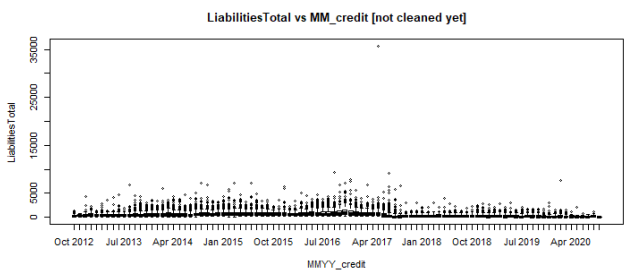
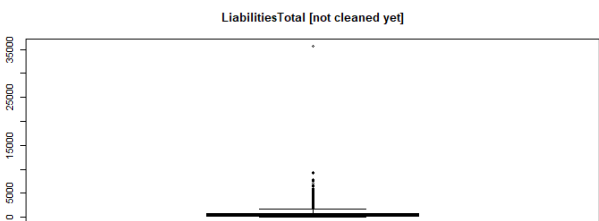
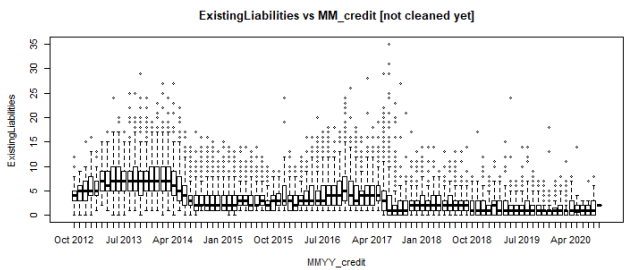
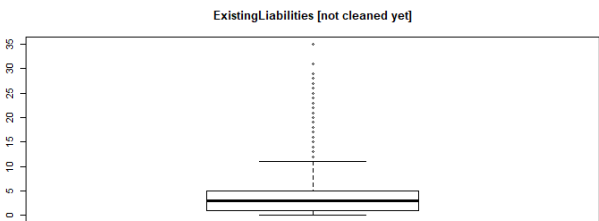
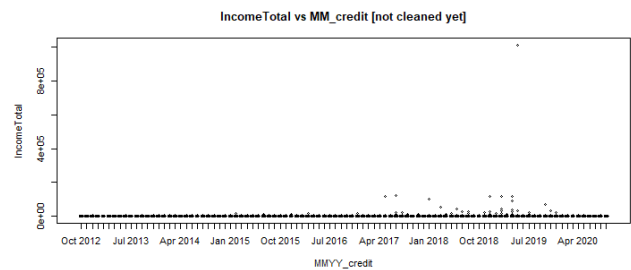
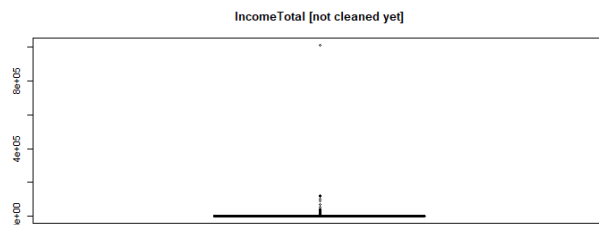
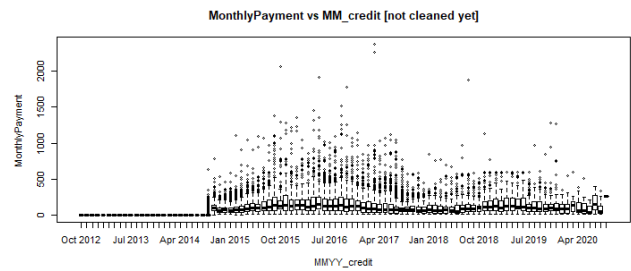
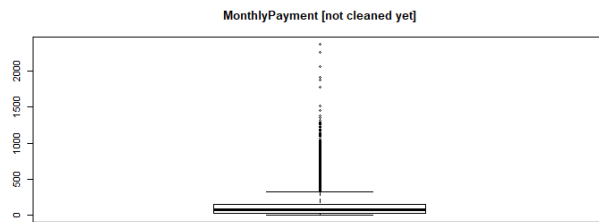
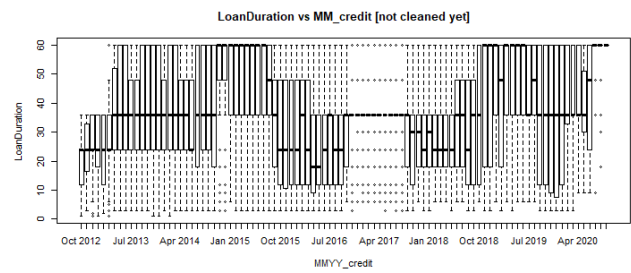
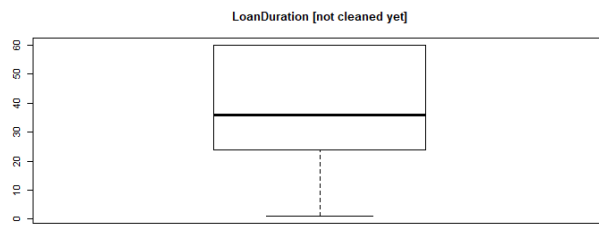


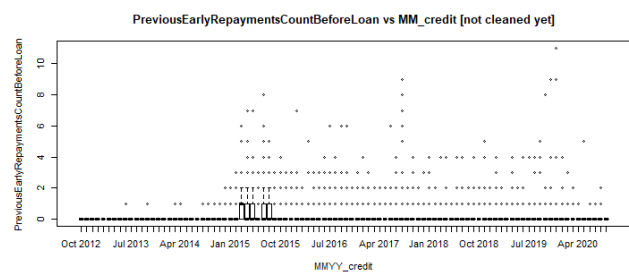
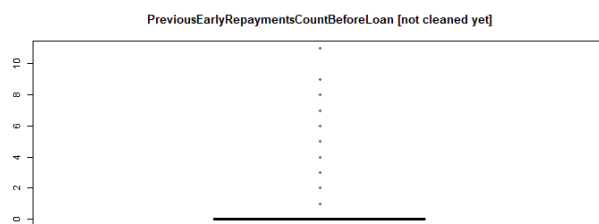
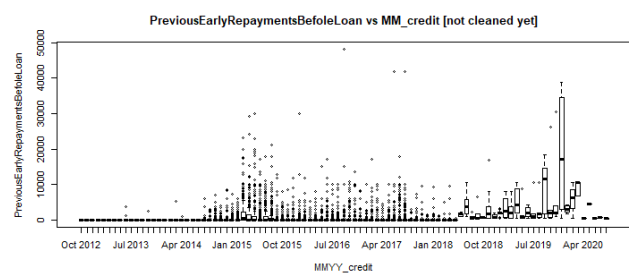
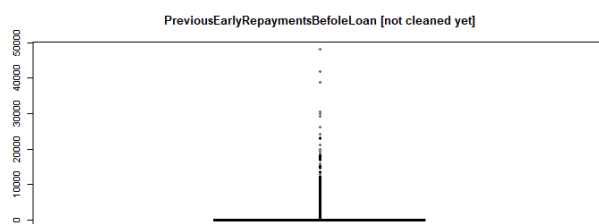
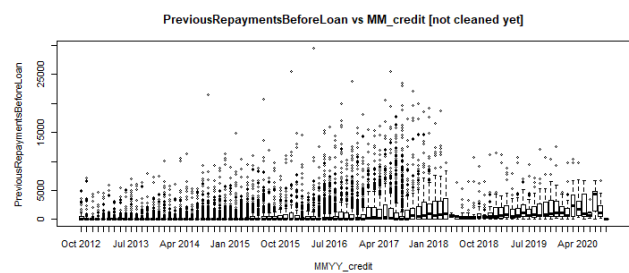
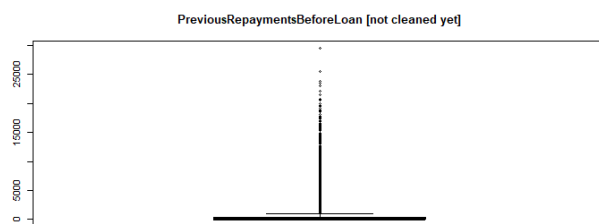
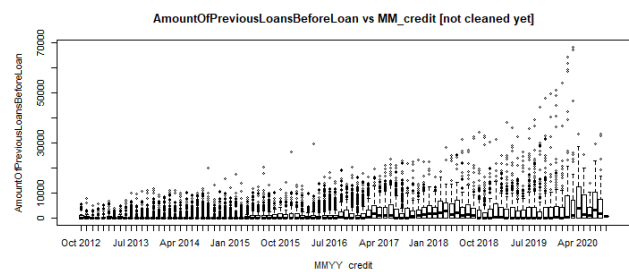
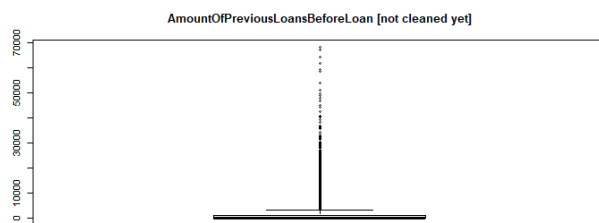
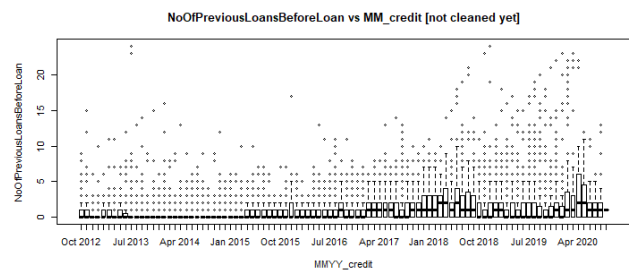
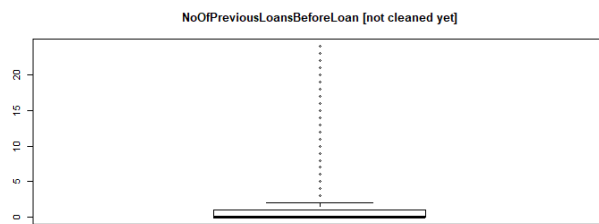
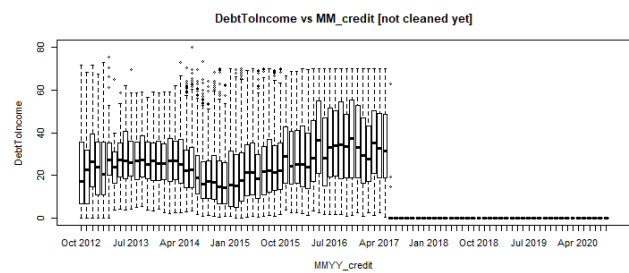
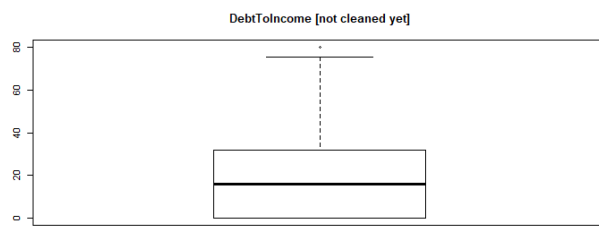
Variable Name	Variable Type
PrincipalWriteOffs	numeric
InterestAndPenalty-WriteOffs	numeric
PrincipalBalance	numeric
InterestAndPenalty-Balance	numeric
NoOfPreviousLoans-BeforeLoan	numeric
AmountOfPrevious-LoansBeforeLoan	numeric
PreviousRepay-mentsBeforeLoan	numeric
PreviousEarlyRepay-mentsBeforeLoan	numeric
PreviousEarlyRepay-mentsCountBeforeLoan	numeric
GracePeriodStart	Date
GracePeriodEnd	Date
NextPaymentDate	Date
NextPaymentNr	numeric
NrOfScheduledPay-ments	numeric
ReScheduledOn	Date
PrincipalDebtServic-ingCost	numeric
InterestAndPenalty-DebtServicingCost	numeric
ActiveLateLastPay-mentCategory	character

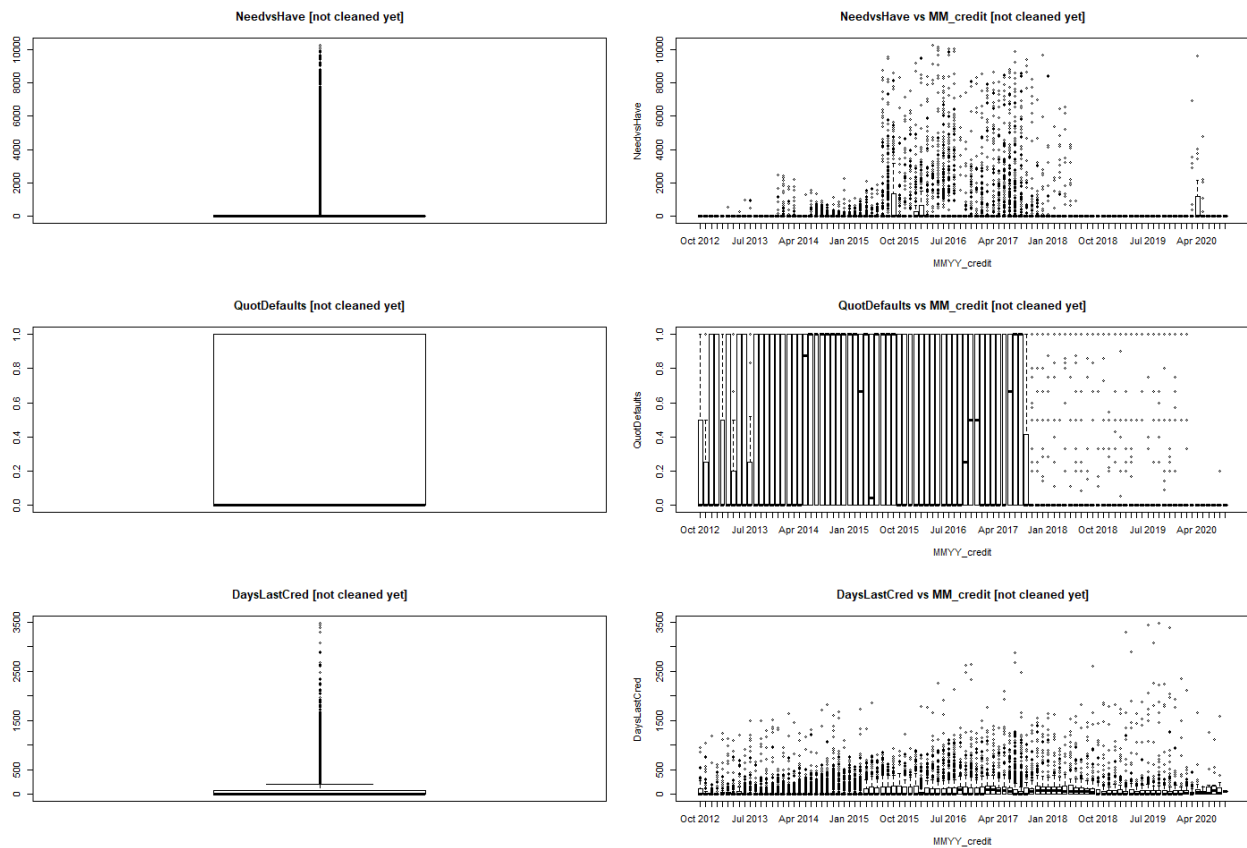
## 6.2 Raw Data – Graphical overview of independent variables [not cleaned yet]



Be aware: This document and the information contained within does not substitute financial advice from professionals.







### 6.3 Raw Data – Table of Correlation between BidsManual and Amount

MMYY_credit	Correlation_BidsManual_Amount	Mean_BidsManual	Mean_Amount
Feb 2009	NA	322.753900	322.7544
Mar 2009	1.00000000	123.936765	123.9360
Apr 2009	1.00000000	107.870378	107.8700
May 2009	1.00000000	161.289738	161.2894
Jun 2009	0.99999999	188.539336	188.5387
Jul 2009	0.99999998	220.924425	220.9252
Aug 2009	0.99999999	232.020761	232.0209
Sep 2009	0.99999998	224.580652	224.5773
Oct 2009	0.99999998	229.926033	229.9260
Nov 2009	0.99999999	198.774467	198.7741
Dec 2009	1.00000000	212.388495	212.3909
Jan 2010	0.99999999	213.851714	213.8526
Feb 2010	0.99999999	220.850230	220.8463
Mar 2010	0.99999930	249.987751	250.0120
Apr 2010	1.00000000	293.703093	293.7018
May 2010	1.00000000	345.786165	345.7870
Jun 2010	0.99999999	365.411164	365.4106
Jul 2010	0.99999997	301.898454	301.9140
Aug 2010	0.99999999	324.262851	324.2695

MMYY_credit	Correlation_BidsManual_Amount	Mean_BidsManual	Mean_Amount
Sep 2010	0.99999998	410.047801	410.0559
Oct 2010	0.99999997	596.643622	596.6475
Nov 2010	0.99999997	513.334838	513.3533
Dec 2010	0.99999998	634.378231	634.3941
Jan 2011	1.00000000	889.500000	889.5000
Feb 2011	1.00000000	798.125000	798.1250
Mar 2011	1.00000000	1081.574074	1081.5741
Apr 2011	1.00000000	1020.729167	1020.7292
May 2011	1.00000000	1105.434783	1105.4348
Jun 2011	1.00000000	910.100000	910.1000
Jul 2011	1.00000000	1206.034483	1206.0345
Aug 2011	1.00000000	1070.526316	1070.5263
Sep 2011	1.00000000	1252.575758	1252.5758
Oct 2011	1.00000000	1316.666667	1316.6667
Nov 2011	1.00000000	1820.322581	1820.3226
Dec 2011	1.00000000	1013.695652	1013.6957
Jan 2012	1.00000000	1069.000000	1069.0000
Feb 2012	1.00000000	1060.576923	1060.5769
Mar 2012	1.00000000	1798.000000	1798.0000
Apr 2012	1.00000000	1481.612903	1481.6129
May 2012	1.00000000	2208.076923	2208.0769
Jun 2012	1.00000000	1921.944444	1921.9444
Jul 2012	1.00000000	1278.064516	1278.0645
Aug 2012	1.00000000	1176.363636	1176.3636
Sep 2012	1.00000000	1560.000000	1560.0000
Oct 2012	0.78591721	699.051724	1493.1034
Nov 2012	After Sep 2012 no correlation coefficient of 1 or close to 1 is achieved anymore	...	...
...	...	...	...

## 6.4 Analysis Data – Graphical overview of independent variables [cleaned]

