Credit Risk Modelling

Luan Cézari Maria

09/06/2020

OVERVIEW

The present work is the conclusion project of Harvardx Profissional Certificate in Data Science. The goal of project is build a model that predict if a customer will or not paid a loan through a bunch of feature. The raw dataset used contain 100514 individual observations and 19 columns been one the target variable and 18 the features. To build this, I first preprocess the data in order to deal with duplicated rows, redundant columns, missing data and other minors errors. Also, in preprocess phase, binarize two variables turning it from numerical to yes/no. After, I perform a analysis of each variable distribution and the relation of some of them. At end, I model the wrangled data using two methods, logistic regression and random forest, comment the obtained results and define future works.

ANALYSIS

Data exploration and cleaning

At beginning we can look at the dataset dimension, first rows and columns types in order to obtain a first impression of it.

Dataset dimension

x
100514
19

First rows

Loan ID	Customer ID	Loan Status	Current Loan Amount	Term
14dd8831-6af5-400b-83ec-68e61888a048	981165ec-3274-42f5-a3b4-d104041a9ca9	Fully Paid	445412	Short Term
4771cc26-131a-45db-b5aa-537ea4ba5342	2de017a3-2e01-49cb-a581-08169e83be29	Fully Paid	262328	Short Term
4eed4e6a-aa2f-4c91-8651-ce984ee8fb26	5efb2b2b-bf11-4dfd-a572-3761a2694725	Fully Paid	99999999	Short Term
77598f7b-32e7-4e3b-a6e5-06ba0d98fe8a	e777faab-98ae-45af-9a86-7ce5b33b1011	Fully Paid	347666	Long Term
d4062e70-befa-4995-8643-a0de73938182	81536ad9-5ccf-4eb8-befb-47a4d608658e	Fully Paid	176220	Short Term
89d8cb0c-e5c2-4f54-b056-48a645c543dd	4ffe99d3-7f2a-44db-afc1-40943f1f9750	Charged Off	206602	Short Term

Credit Score	Annual Income	Years in current job	Home Ownership	Purpose
709	1167493	8 years	Home Mortgage	Home Improvements
NA	NA	10+ years	Home Mortgage	Debt Consolidation
741	2231892	8 years	Own Home	Debt Consolidation
721	806949	3 years	Own Home	Debt Consolidation
NA	NA	5 years	Rent	Debt Consolidation
7290	896857	10+ years	Home Mortgage	Debt Consolidation

Monthly Debt	Years of Credit History	Months since last delinquent	Number of Open Accounts	Number of Credit Problems
5214.74	17.2	NA	6	1
33295.98	21.1	8	35	0
29200.53	14.9	29	18	1
8741.90	12.0	NA	9	0
20639.70	6.1	NA	15	0
16367.74	17.3	NA	6	0

Current Credit Balance	Maximum Open Credit	Bankruptcies	Tax Liens
228190	416746	1	0
229976	850784	0	0
297996	750090	0	0
256329	386958	0	0
253460	427174	0	0
215308	272448	0	0

Columns types

key	Class
Loan ID	character
Customer ID	character
Loan Status	character
Term	character
Years in current job	character
Home Ownership	character
Purpose	character
Current Loan Amount	numeric
Credit Score	numeric
Annual Income	numeric
Monthly Debt	numeric
Years of Credit History	numeric
Months since last delinquent	numeric
Number of Open Accounts	numeric
Number of Credit Problems	numeric
Current Credit Balance	numeric
Maximum Open Credit	numeric
Bankruptcies	numeric
Tax Liens	numeric

Here we can percept that the raw dataset contains 100514 individual observations with 19 columns, been one the target variable and 18 features. Of these, 12 are numeric variables and 6 are character. Two of the character features, Loan ID and Customer ID, don't have any predictive power and so should be removed. At next, we explore the missing data:

key	NAs	NAs Proportion
Months since last delinquent	53655	0.5338062
Credit Score	19668	0.1956742
Annual Income	19668	0.1956742
Bankruptcies	718	0.0071433
Tax Liens	524	0.0052132
Maximum Open Credit	516	0.0051336
Loan ID	514	0.0051137
Customer ID	514	0.0051137
Loan Status	514	0.0051137
Current Loan Amount	514	0.0051137
Term	514	0.0051137
Years in current job	514	0.0051137
Home Ownership	514	0.0051137
Purpose	514	0.0051137
Monthly Debt	514	0.0051137
Years of Credit History	514	0.0051137
Number of Open Accounts	514	0.0051137
Number of Credit Problems	514	0.0051137
Current Credit Balance	514	0.0051137

Here we can see that there is a variable with more than 50% of the missing data. As it is a very large proportion, the best way to deal with this is to remove the column completely. We can also see that the other two lines with significant NAs have the same number of them and, therefore, it is possible that them comes from the same observations. If confirmed, we can continue removing the lines.

Are the observations the same?
TRUE

If we look closely at the distribution of Credit Score

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##     585     705     724    1076     741    7510   19668 
```

number of credit score bigger than 850
4551

we see that there are 4551 credit scores greater than 850. As the maximum FICO score is 850, this is impossible and therefore we can interpret this as a human error. To resolve this, I divide the credit score above 850 to 10.

Next, I check if there's observations with Current Loan Amount bigger than Maximum Open Credit. As this is impossible, due to the fact that the current loan amount is a credit that has been opened, we can interpret this as an error. As we can't recover the correct information the best way to proceed is remove the respective rows.

```r
# Checking if there's observations with Current Loan Amount bigger than Maximum Open Credit
raw_data %>%
  summarize(n = sum(`Current Loan Amount` > `Maximum Open Credit`, na.rm = TRUE)) %>%
 kable() %>%
 kable_styling(latex_options = c("striped", "bordered"))
```

n
30935

I will now deal with bankruptcies, tax burdens and number of appeals from credit problems. I start by looking at the distribution of these variables.

	0	1	2	3	4	5	6	7	8	9	10	11	12	15
Credit Problems	86035.00	12077.000	1299.000	378.000	125.000	49	17	8	4	2	2	2	1	1
proportion	0.86	0.121	0.013	0.004	0.001	0	0	0	0	0	0	0	0	0

	0	1	2	3	4	5	6	7	9	10	11	15
Tax Liens	98062.000	1343.000	374.000	111.000	58.000	16	12	7	3	1	2	1
proportion	0.981	0.013	0.004	0.001	0.001	0	0	0	0	0	0	0

	0	1	2	3	4	5	6	7
Bankruptcies	88774.00	10475.000	417.000	93.000	27	7	2	1
proportion	0.89	0.105	0.004	0.001	0	0	0	0

Here we can see that there are few observations with a value greater than 2 in any of these variables. We can then consider the binarization of this, making the variables yes / no instead of numeric. We can also see that bankruptcies are credit problems and there are few individuals with credit problems who have not gone bankrupt. Looking at this relationship numerically, as we do below, I can conclude that the variable Bankruptcies is redundant and can be removed with almost no loss of predictability.

n	proportion
2935	0.02935

Next I look for the unique data in each column:

key	uniques
Loan ID	81999
Customer ID	81999
Monthly Debt	65765
Maximum Open Credit	44596
Annual Income	36174
Current Credit Balance	32730
Current Loan Amount	22004
Years of Credit History	506
Credit Score	324
Months since last delinquent	116
Number of Open Accounts	51
Purpose	16
Number of Credit Problems	14
Years in current job	12
Tax Liens	12
Bankruptcies	8
Home Ownership	4
Loan Status	2
Term	2

Here it's possible to perception that the Loan ID number is less than the number of observations. This indicates that there are duplicate lines and they must be removed. Now, we look at the unique values of categorical features, excluding Loan ID and Customer ID:

```
## $`Loan Status`
## [1] "Fully Paid"  "Charged Off" NA
##
## $Term
## [1] "Short Term" "Long Term"  NA
##
## $`Home Ownership`
## [1] "Home Mortgage" "Own Home"      "Rent"          "HaveMortgage"
## [5] NA
##
## $`Years in current job`
##  [1] "8 years"   "10+ years" "3 years"   "5 years"   "< 1 year"  "2 years"
##  [7] "4 years"   "9 years"   "7 years"   "1 year"    "n/a"       "6 years"
## [13] NA
##
## $Purpose
##  [1] "Home Improvements"    "Debt Consolidation"   "Buy House"
##  [4] "other"                "Business Loan"        "Buy a Car"
##  [7] "major_purchase"       "Take a Trip"          "Other"
## [10] "small_business"       "Medical Bills"        "wedding"
## [13] "vacation"             "Educational Expenses" "moving"
## [16] "renewable_energy"     NA
```

Here I can see many problems that need to be solved.
Initially, "Residential mortgage" and "Termortgage" in "Domestic property" mean the same and, therefore, one must be converted into the other.
In "Years of current work", there is a character value "n / a" that must be converted to NA. In "Purpose", there are two different "others", one with the lowercase O and the other with the uppercase, both need to be combined into one. In addition, there is a "commercial loan" and a "small_business". Since "small

business loan" is a type of "commercial loan" and we cannot confirm whether or not there is a small business in"commercial loan", it is good to convert both into a single value.
I finish by removing any remaining NA lines and converting character columns into factor.

Now, after apply all of these transformations, I evaluate the resulting dataset.

key	Class	Uniques	NAs
Monthly Debt	numeric	39203	0
Maximum Open Credit	numeric	31243	0
Annual Income	numeric	27487	0
Current Credit Balance	numeric	25888	0
Current Loan Amount	numeric	17785	0
Years of Credit History	numeric	476	0
Credit Score	numeric	167	0
Number of Open Accounts	numeric	50	0
Purpose	factor	14	0
Years in current job	factor	11	0
Home Ownership	factor	3	0
Loan Status	factor	2	0
Term	factor	2	0
Tax Liens	factor	2	0
Historic of Credit Problems	factor	2	0

key	value
none credit score with decimal	TRUE

```
## [1] 44112    15
```

All looks to be correct. The wrangled data has 44112 observations and 15 columns been 14 features.
After all these transformations, we need to conclude this stage of the process by creating the train and test set through the wrangled data, as shown below:

This process is important because it helps to avoid excessive adjustments in the later modeling phase. I choose to divide it as 80% of the data in the training set and 20% of the data in the test set, because using more data in the train set reduces the variability of the model result and I use 80% of the train set instead of 90% because it makes the training stage computationally less intensive.

Data visualization

In this section, we will visually explore the distribution of variables and the relationship between some of them in the training set. We will start by visualizing the distribution of the target variable, then we will analyze the distribution and the proportion of the discrete variables, the continuous variables and, finally, we will analyze the relationship between them.

Target variable

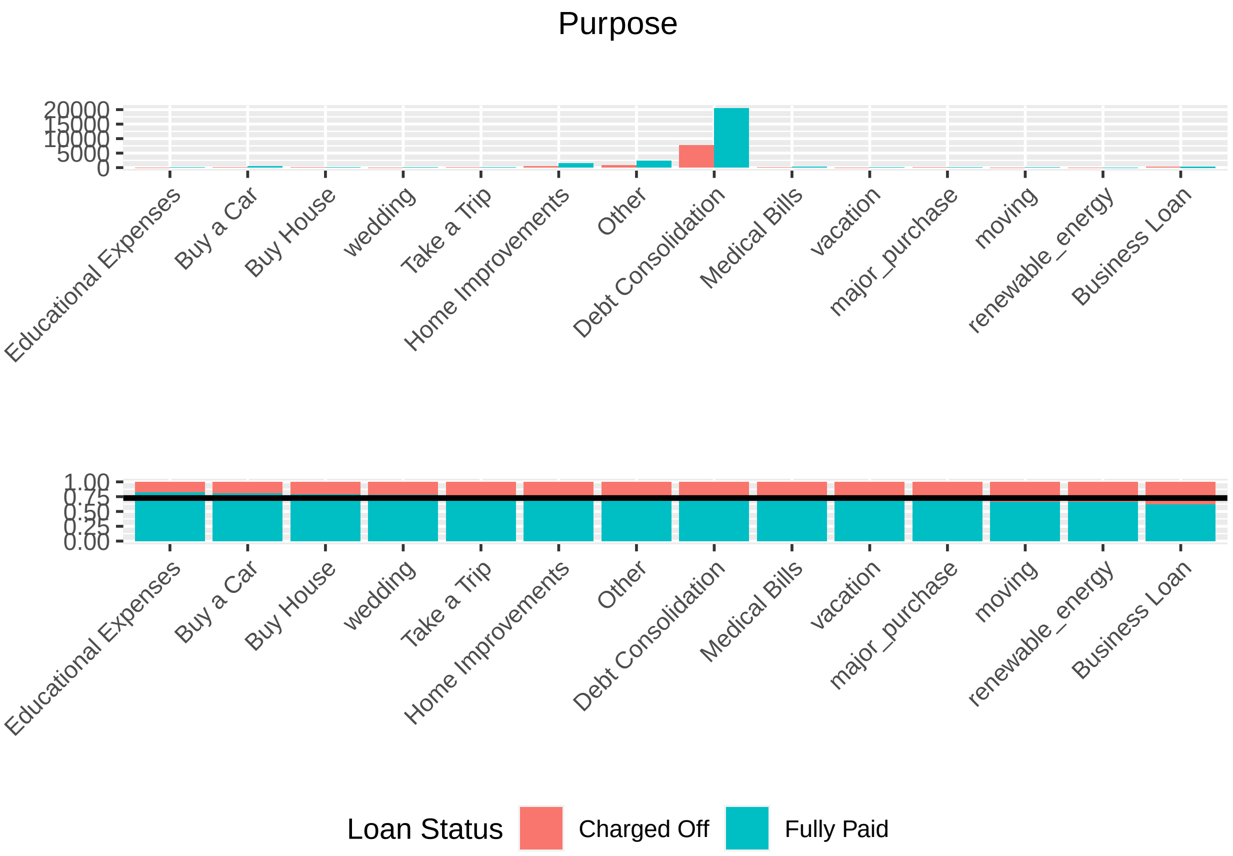
```
CrossTable(train_set$`Loan Status`)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  35289
##
##
##             | Charged Off |  Fully Paid |
##             |-------------|-------------|
##             |        9626 |       25663 |
##             |       0.273 |       0.727 |
##             |-------------|-------------|
##
##
##
##
```

Here we can see that there is a much greater distribution of “Fully Paid” than “Charged Off”. This difference can lead to inaccurate results during the training phase and will therefore be taken into account.

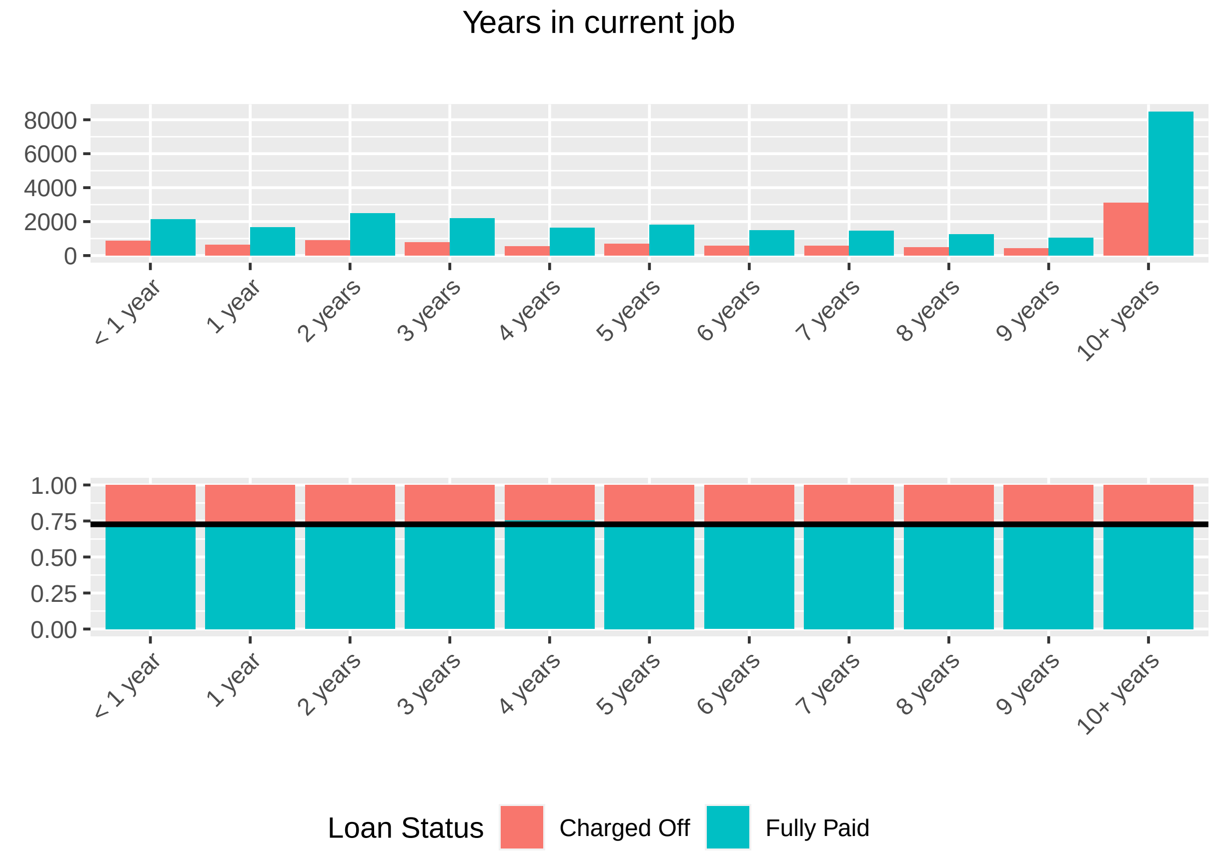
Categorical features distribution

Purpose



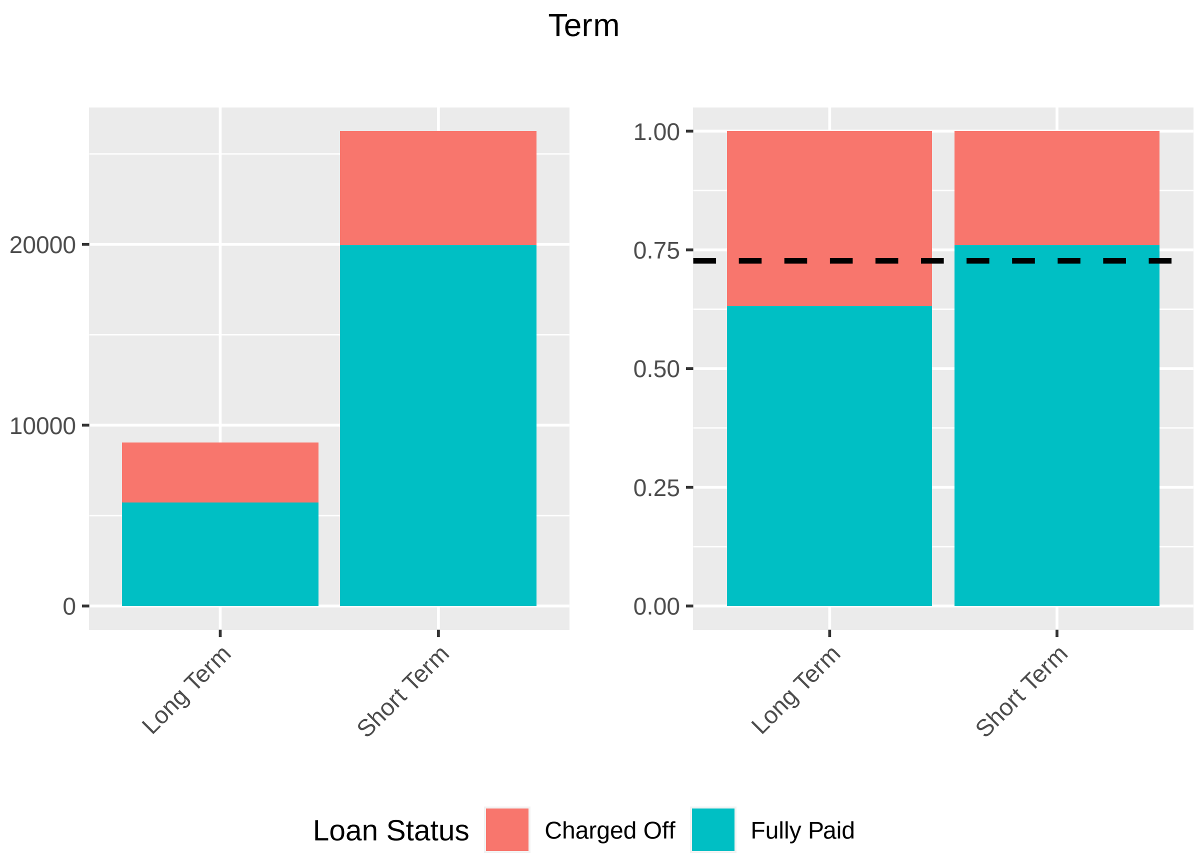
Here we can see that almost all of the training set loans were for debt consolidation. In terms of proportion to status, some purposes such as business loans and moving tend to have a higher proportion of Charged Off loans than the whole as a whole while others such as educational expenses and buy a car tend to a greater proportion of Fully Paid than that reference.

Years in current job



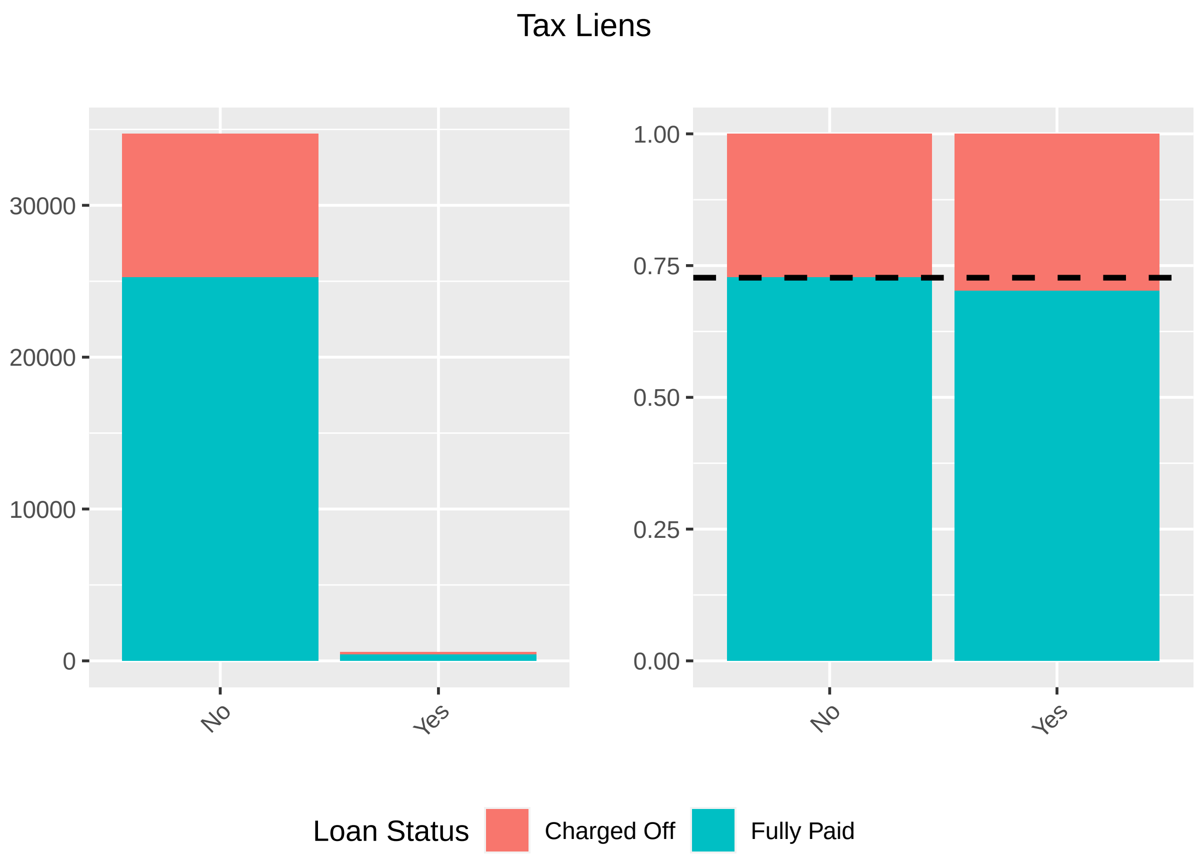
Here we can see that the vast majority of loans were made by people aged 10 or over in their current job. The proportion of Loan Status between the different periods was not significant.

Term



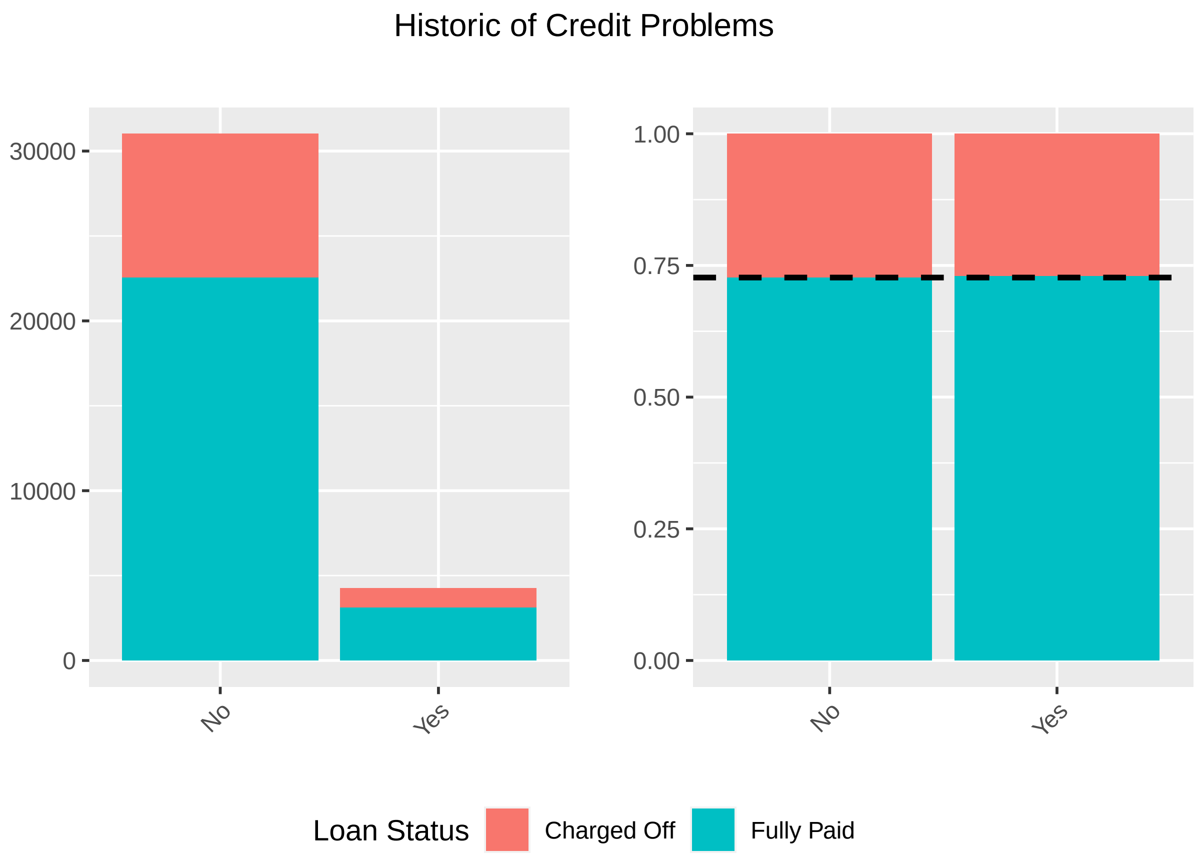
Here we can see that the vast majority of loans were made in the short term, being proportionately more paid than those made in the long term.

Tax Liens



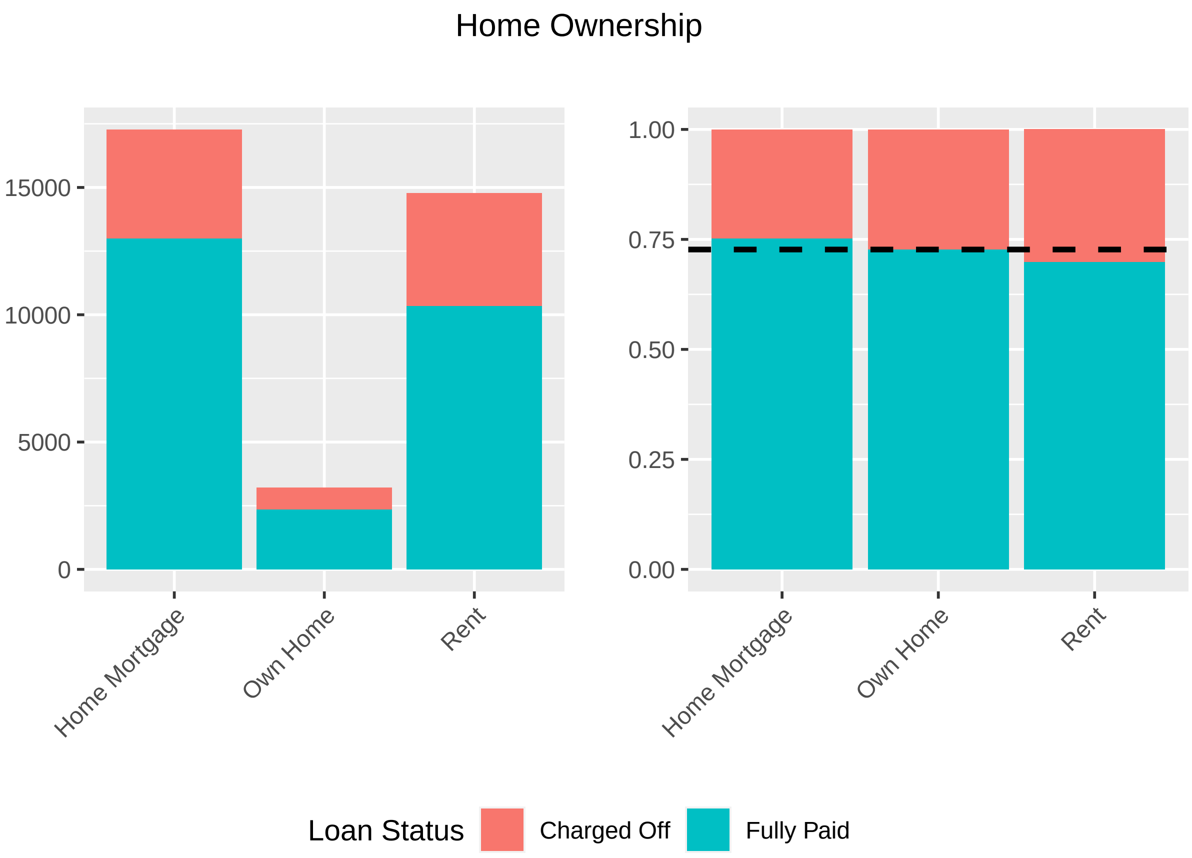
Only a small, almost insignificant group of people had tax liens problems. These proportionally less honored the loan contracts.

Historic of Credit Problems



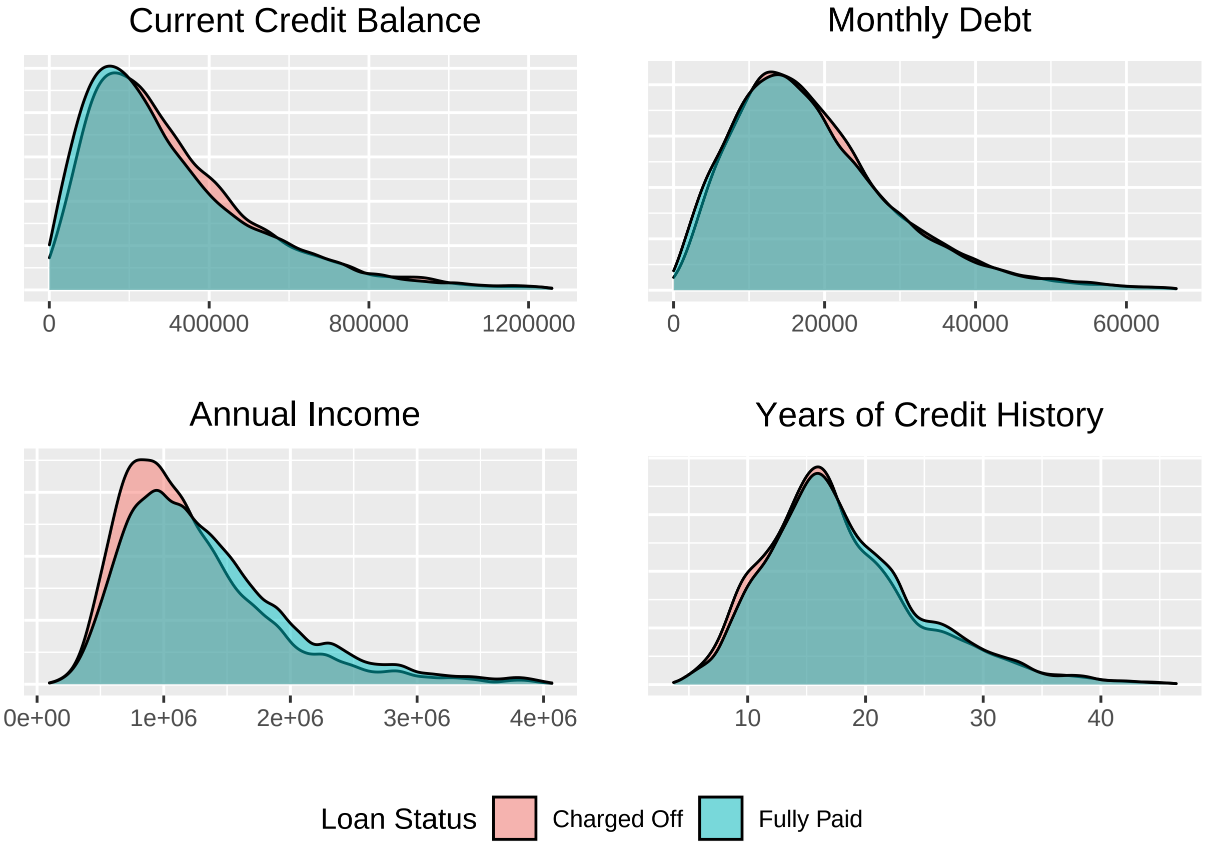
Most borrowers have never had credit problems. The proportion of individuals who have honored loan agreements and have a history of credit problems is statistically the same as those who do not have the same history.

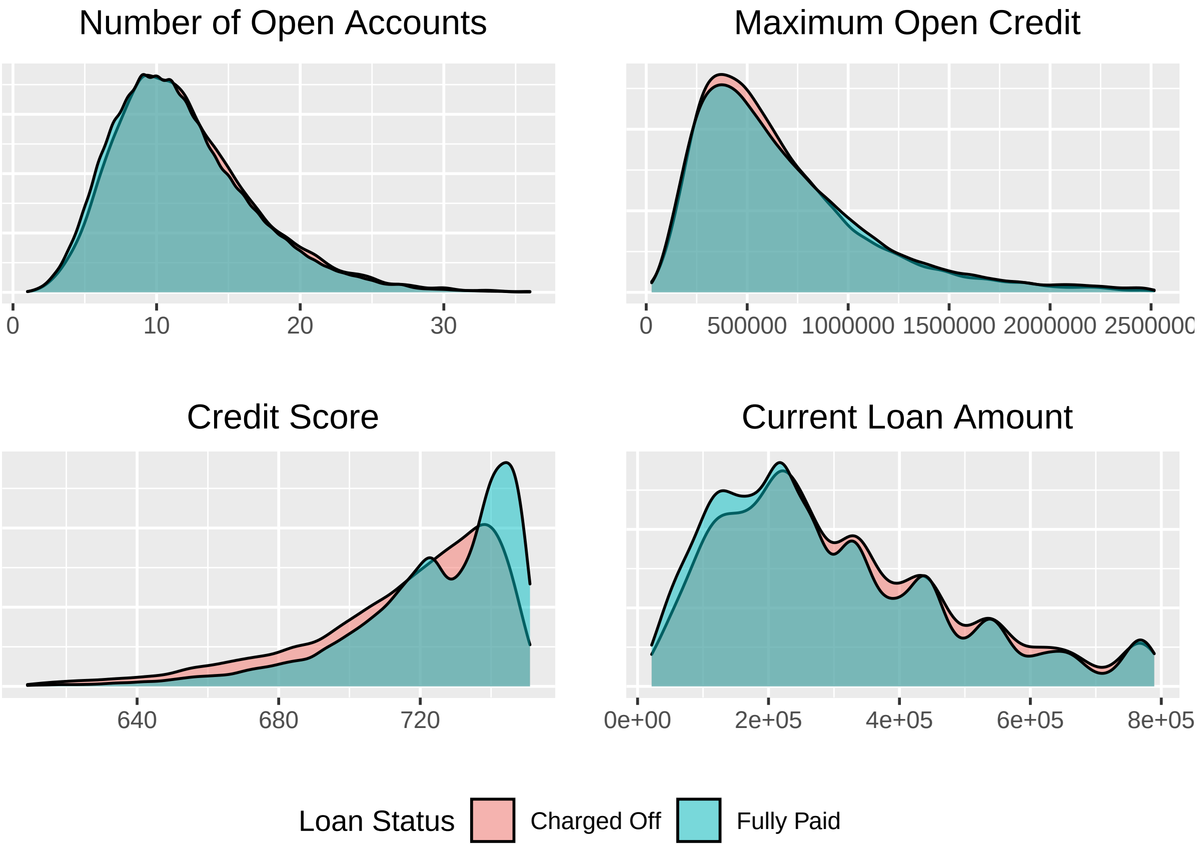
Home Ownership



The smallest part of the borrowers pays a mortgage while the smallest part of them own their own home. Proportionally, mortgage-paying customers tend to pay more on their loans than the general average, while those who live on rent tend to pay less.

Continuous features distribution



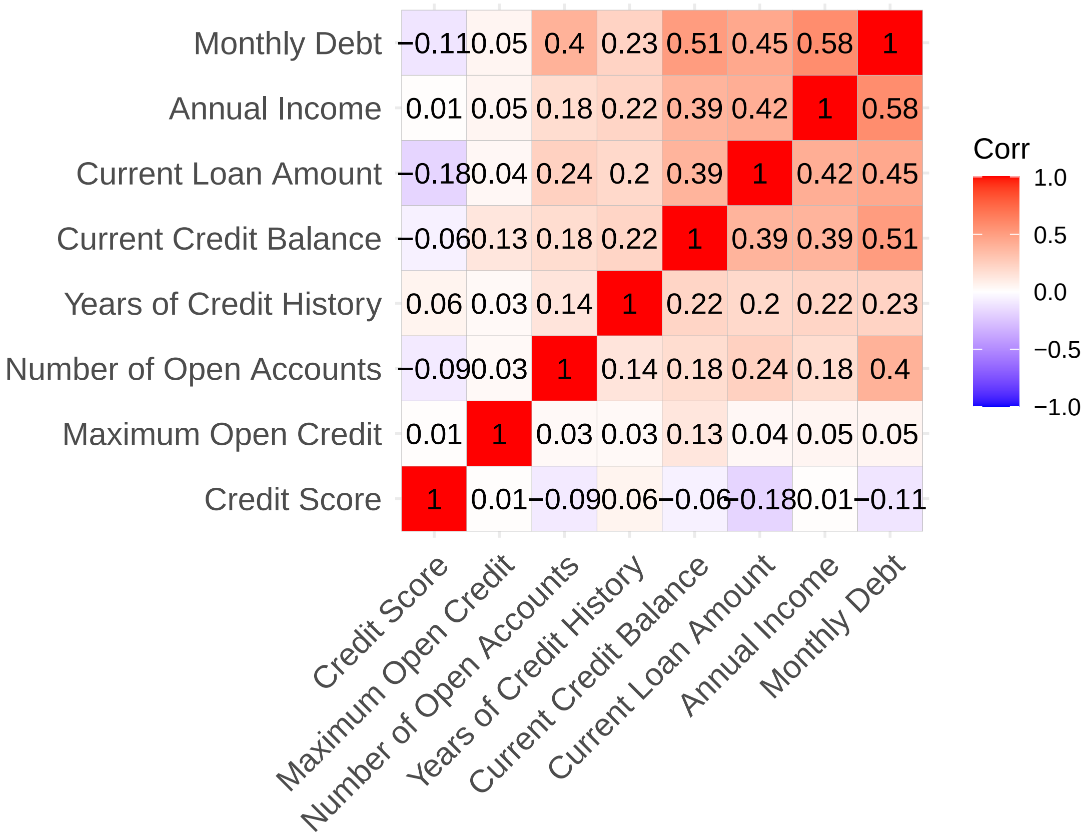


Here we can see that for most continuous variables there is no significant difference in the distribution of those who paid the loans and those who did not. Only the annual income, credit score and current loan amount differ significantly.

The proportion of "Fully Paid" becomes greater after a certain point in the variables annual income, years of credit history, maximum open credit and credit score, with the proviso that in the case of credit score there are two local maximum points where the distribution of "Fully Paid" is greater with a slope in the middle where the proportion of "Charged Off" stands out.

Relation between variables

Correlation matrix of continuous variables



Here it is noticeable that the vast majority of variables have a positive correlation with each other, with the exception of the credit score variable, which has a negative correlation with the vast majority of the others. The variables monthly debt and annual income are those that tend to have the highest correlation with the others, but none of them are correlated enough to influence negatively the efficiency of the models.

Cross-table between historic of credit problems and tax liens

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  35289
##
##
##                                         | train_set$`Tax Liens`
## train_set$`Historic of Credit Problems` |        No |       Yes | Row Total |
## ----------------------------------------|-----------|-----------|-----------|
##                                      No |     31020 |         0 |     31020 |
##                                         |     1.000 |     0.000 |     0.879 |
##                                         |     0.894 |     0.000 |           |
##                                         |     0.879 |     0.000 |           |
## ----------------------------------------|-----------|-----------|-----------|
##                                     Yes |      3689 |       580 |      4269 |
##                                         |     0.864 |     0.136 |     0.121 |
##                                         |     0.106 |     1.000 |           |
##                                         |     0.105 |     0.016 |           |
## ----------------------------------------|-----------|-----------|-----------|
##                            Column Total |     34709 |       580 |     35289 |
##                                         |     0.984 |     0.016 |           |
## ----------------------------------------|-----------|-----------|-----------|
##
##
```

It is notable that there are no individuals who have had tax liens problems and have had no history of credit problems. It is also notable that very few have had tax liens problems and a history of credit problems at the same time.

Modeling

Adjusting the target variable

Before run the predict models we need to do some minor adjusts in target variable. First, we need add a underline between the two words of each unique value, obtaining the levels "Fully_Paid" and "Chaged Off". This stage is necessary due to especifications of the functions used. Then, we need to change the order of factors putting the "Fully_Paid" as the first level, so it will be automaticaly setted as the positive class in the models.

Choosing parameter evaluation metric

As has been seen in data visualization there's a high prevalence of "Fully Paid" over "Charged Off". This characteristic of target variable can lead to a model with big accuracy even with a lack of capacity to distinguish between two possible outcomes. To overcome this I choose to use ballance accuracy instead of accuracy as metric to fit the models.

```
# Create balanced accuracy function
baSummary <- function(data, lev = NULL, model = NULL){
  out <- (sensitivity(data$pred, data$obs)+specificity(data$pred, data$obs))/2
  c(balancedAccuracy = out)
}
```

Logistic regression model 1

The logistic regression is one of the most common classification model used due to his simplicity and low computational intensity. We will try to model a logistic regression using all avaliable features.
We define the outcome Y as 1 for "Fully Paid" and for a vector of predictors $\boldsymbol{X}$ we estimate the conditional probability $Pr(Y = 1|X_n = x_n) = \beta_0 + \sum_{n=1}^{k} X_n\beta_n$. Due to the fact that this linear function can take values larger than 1 we need to use the logistic transformation $g(p) = log\frac{p}{1-p}$ obtaining the final model $g\{Pr(Y = 1|X_n = x_n)\} = \beta_0 + \sum_{n=1}^{k} X_n\beta_n$. We use, then, the maximum likelihood estimate to estimate the β.

```
glm_fit <- train(`Loan Status` ~ .,
                 data = train_set,
                 method = "glm",
                 family = "binomial")
```

Doing so we obtain te results

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5661  -0.8130  -0.6961   1.2492   5.4077
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             5.134e+00  3.684e-01  13.934  < 2e-16
## `\\`Current Loan Amount\\``             4.697e-07  9.167e-08   5.124 2.99e-07
## `TermShort Term`                       -3.881e-01  3.300e-02 -11.761  < 2e-16
## `\\`Credit Score\\``                   -7.246e-03  5.136e-04 -14.108  < 2e-16
## `\\`Annual Income\\``                  -4.864e-07  2.582e-08 -18.839  < 2e-16
## `\\`Years in current job\\`1 year`     -3.743e-02  6.290e-02  -0.595 0.551851
## `\\`Years in current job\\`10+ years`  -4.195e-02  4.776e-02  -0.878 0.379746
## `\\`Years in current job\\`2 years`    -8.108e-02  5.737e-02  -1.413 0.157577
## `\\`Years in current job\\`3 years`    -1.082e-01  5.941e-02  -1.822 0.068520
## `\\`Years in current job\\`4 years`    -1.855e-01  6.552e-02  -2.832 0.004630
## `\\`Years in current job\\`5 years`    -1.847e-02  6.165e-02  -0.300 0.764539
## `\\`Years in current job\\`6 years`     7.726e-05  6.493e-02   0.001 0.999051
## `\\`Years in current job\\`7 years`    -6.393e-03  6.545e-02  -0.098 0.922186
## `\\`Years in current job\\`8 years`    -1.257e-02  6.877e-02  -0.183 0.854904
## `\\`Years in current job\\`9 years`     9.142e-02  7.206e-02   1.269 0.204561
## `\\`Home Ownership\\`Own Home`          8.151e-02  4.472e-02   1.823 0.068348
## `\\`Home Ownership\\`Rent`              2.600e-01  2.801e-02   9.282  < 2e-16
## `PurposeBuy a Car`                     -8.780e-01  1.507e-01  -5.828 5.60e-09
## `PurposeBuy House`                     -9.379e-01  2.009e-01  -4.668 3.05e-06
## `PurposeDebt Consolidation`            -5.517e-01  9.063e-02  -6.087 1.15e-09
## `PurposeEducational Expenses`          -9.643e-01  5.111e-01  -1.887 0.059227
## `PurposeHome Improvements`             -4.571e-01  1.056e-01  -4.331 1.48e-05
## Purposemajor_purchase                  -2.373e-01  2.278e-01  -1.042 0.297640
## `PurposeMedical Bills`                 -3.933e-01  1.452e-01  -2.710 0.006735
## Purposemoving                          -3.209e-01  2.964e-01  -1.083 0.278969
## PurposeOther                           -5.354e-01  9.818e-02  -5.453 4.95e-08
## Purposerenewable_energy                -2.999e-01  1.323e+00  -0.227 0.820611
## `PurposeTake a Trip`                   -6.454e-01  1.885e-01  -3.424 0.000617
## Purposevacation                        -3.759e-01  3.945e-01  -0.953 0.340676
## Purposewedding                         -7.280e-01  4.174e-01  -1.744 0.081157
## `\\`Monthly Debt\\``                    1.354e-05  1.571e-06   8.617  < 2e-16
## `\\`Years of Credit History\\``        -2.398e-03  1.957e-03  -1.226 0.220332
## `\\`Number of Open Accounts\\``         8.841e-03  2.651e-03   3.335 0.000854
## `\\`Current Credit Balance\\``          7.791e-08  4.495e-08   1.733 0.083052
## `\\`Maximum Open Credit\\``            -1.269e-08  1.157e-08  -1.096 0.272931
## `\\`Tax Liens\\`Yes`                    2.245e-01  1.014e-01   2.213 0.026871
## `\\`Historic of Credit Problems\\`Yes` -3.550e-02  4.103e-02  -0.865 0.386901
##
## (Intercept)                            ***
## `\\`Current Loan Amount\\``            ***
## `TermShort Term`                       ***
## `\\`Credit Score\\``                   ***
## `\\`Annual Income\\``                  ***
## `\\`Years in current job\\`1 year`
## `\\`Years in current job\\`10+ years`
## `\\`Years in current job\\`2 years`
## `\\`Years in current job\\`3 years`      .
## `\\`Years in current job\\`4 years`      **
## `\\`Years in current job\\`5 years`
## `\\`Years in current job\\`6 years`
## `\\`Years in current job\\`7 years`
## `\\`Years in current job\\`8 years`
## `\\`Years in current job\\`9 years`
## `\\`Home Ownership\\`Own Home`          .
## `\\`Home Ownership\\`Rent`              ***
## `PurposeBuy a Car`                     ***
## `PurposeBuy House`                     ***
## `PurposeDebt Consolidation`            ***
## `PurposeEducational Expenses`          .
## `PurposeHome Improvements`             ***
## Purposemajor_purchase
## `PurposeMedical Bills`                 **
## Purposemoving
## PurposeOther                           ***
## Purposerenewable_energy
## `PurposeTake a Trip`                   ***
## Purposevacation
## Purposewedding                         .
## `\\`Monthly Debt\\``                   ***
## `\\`Years of Credit History\\``
## `\\`Number of Open Accounts\\``        ***
## `\\`Current Credit Balance\\``         .
## `\\`Maximum Open Credit\\``
## `\\`Tax Liens\\`Yes`                   *
## `\\`Historic of Credit Problems\\`Yes`
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 41359  on 35288  degrees of freedom
## Residual deviance: 39781  on 35252  degrees of freedom
## AIC: 39855
##
## Number of Fisher Scoring iterations: 7
```

Logistic regression model 2

We can try to improve this model removing the features with p-value bigger than 0.01 obtaining the model below:

```
glm_fit2 <- train(`Loan Status` ~ `Current Loan Amount` + `Term` + `Credit Score` + `Annual Income` + `
                  data = train_set,
                  method = "glm",
                  family = "binomial")
```

```
##
## Call:
## NULL
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5721  -0.8136  -0.6973   1.2499   5.4121
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      5.125e+00  3.628e-01  14.126  < 2e-16 ***
## `\\`Current Loan Amount\\``      4.897e-07  8.990e-08   5.448 5.11e-08 ***
## `TermShort Term`                -3.858e-01  3.294e-02 -11.712  < 2e-16 ***
## `\\`Credit Score\\``            -7.364e-03  5.066e-04 -14.537  < 2e-16 ***
## `\\`Annual Income\\``           -4.864e-07  2.563e-08 -18.976  < 2e-16 ***
## `PurposeBuy a Car`              -8.721e-01  1.506e-01  -5.792 6.97e-09 ***
## `PurposeBuy House`              -9.368e-01  2.007e-01  -4.667 3.06e-06 ***
## `PurposeDebt Consolidation`     -5.418e-01  9.044e-02  -5.991 2.09e-09 ***
## `PurposeEducational Expenses`   -9.374e-01  5.107e-01  -1.835 0.066433 .
## `PurposeHome Improvements`      -4.571e-01  1.054e-01  -4.336 1.45e-05 ***
## Purposemajor_purchase           -2.280e-01  2.276e-01  -1.002 0.316536
## `PurposeMedical Bills`          -3.908e-01  1.450e-01  -2.695 0.007035 **
## Purposemoving                   -3.073e-01  2.964e-01  -1.037 0.299839
## PurposeOther                    -5.298e-01  9.810e-02  -5.400 6.66e-08 ***
## Purposerenewable_energy         -2.608e-01  1.311e+00  -0.199 0.842334
## `PurposeTake a Trip`            -6.394e-01  1.883e-01  -3.397 0.000682 ***
## Purposevacation                 -3.512e-01  3.942e-01  -0.891 0.372947
## Purposewedding                  -7.193e-01  4.172e-01  -1.724 0.084688 .
## `\\`Monthly Debt\\``             1.433e-05  1.489e-06   9.621  < 2e-16 ***
## `\\`Number of Open Accounts\\``  8.215e-03  2.625e-03   3.130 0.001750 **
## `\\`Home Ownership\\`Own Home`   7.975e-02  4.463e-02   1.787 0.073976 .
## `\\`Home Ownership\\`Rent`       2.588e-01  2.735e-02   9.461  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 41359  on 35288  degrees of freedom
## Residual deviance: 39811  on 35267  degrees of freedom
## AIC: 39855
##
## Number of Fisher Scoring iterations: 4
```

Random Forest Model

The random forest is a supervised learning algorithm. He works by creating a set of decision trees using the bagging method. The benefits of this are that the average of multiple decision trees reduces instability and improves accuracy. Now, I will build a random forest model using all available resources.

```
# Create train control
train.control <- trainControl(method = "cv",
                              number = 5,
                              p = .8,
                              classProbs = TRUE,
                              summaryFunction = baSummary)


# Training model
set.seed(1)
```

```
rf_fit <- train(`Loan Status` ~ .,
                data = train_set,
                method = "rf",
                metric = "balancedAccuracy",
                trControl = train.control,
                tuneGrid = data.frame(mtry = seq(500, 1000, 250)),
                importance = TRUE)
```

RESULTS

Logistic regression model 1

```
## Confusion Matrix and Statistics
##
##              Reference
## Prediction    Fully_Paid Charged_Off
##   Fully_Paid        6314        2288
##   Charged_Off        102         119
##
##                Accuracy : 0.7291
##                  95% CI : (0.7197, 0.7384)
##     No Information Rate : 0.7272
##     P-Value [Acc > NIR] : 0.3472
##
##                   Kappa : 0.0468
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.98410
##             Specificity : 0.04944
##          Pos Pred Value : 0.73402
##          Neg Pred Value : 0.53846
##              Prevalence : 0.72719
##          Detection Rate : 0.71563
##    Detection Prevalence : 0.97495
##       Balanced Accuracy : 0.51677
##
##        'Positive' Class : Fully_Paid
##
```

The logistic regression model 1 clearly failed to differentiate between loans that would potentially be paid in full and loans that would not be. Most of the loans in the training data were predicted to be fully paid and therefore 98.4% of Fully paid loans were predicted correctly while only 4.9% of Charged off were. There may be several causes of this behavior, ranging from the variables considered in the model to the lack of flexibility of the linear regression itself.

Logistic regression model 2

```
## Confusion Matrix and Statistics
##
##              Reference
## Prediction    Fully_Paid Charged_Off
##   Fully_Paid        6314        2291
##   Charged_Off        102         116
```

```
##
##                Accuracy : 0.7288
##                  95% CI : (0.7194, 0.738)
##     No Information Rate : 0.7272
##     P-Value [Acc > NIR] : 0.3741
##
##                   Kappa : 0.0451
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.98410
##             Specificity : 0.04819
##          Pos Pred Value : 0.73376
##          Neg Pred Value : 0.53211
##              Prevalence : 0.72719
##          Detection Rate : 0.71563
##    Detection Prevalence : 0.97529
##       Balanced Accuracy : 0.51615
##
##        'Positive' Class : Fully_Paid
##
```

After removing variables with a p-value greater than 0.01, a worsening of the model's ability to correctly predict the status of loans in training data was observed. In the second model, a sensitivity of 0.984 and a specificity of 0.482 were observed. The difference in performance between the two models, however, was very small and can be easily attributed to chance. After these two models, we can conclude that logistic regression is too rigid to adapt to the data, requiring a more complex approach.

Random forest model

```
## Confusion Matrix and Statistics
##
##              Reference
## Prediction    Fully_Paid Charged_Off
##   Fully_Paid        6138        2174
##   Charged_Off        278         233
##
##                Accuracy : 0.7221
##                  95% CI : (0.7126, 0.7314)
##     No Information Rate : 0.7272
##     P-Value [Acc > NIR] : 0.8615
##
##                   Kappa : 0.0709
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.9567
##             Specificity : 0.0968
##          Pos Pred Value : 0.7385
##          Neg Pred Value : 0.4560
##              Prevalence : 0.7272
##          Detection Rate : 0.6957
##    Detection Prevalence : 0.9421
##       Balanced Accuracy : 0.5267
##
```

```
##        'Positive' Class : Fully_Paid
##
```

Using the random forest approach with all variables we can see an improvement in performance. Although the sensitivity was lower than the logistic regression, reaching 0.9567, the specificity was higher, at 0.097, which is reflected in an improvement in balanced accuracy.

CONCLUSION

In the present work, we analyze a set of loan data and try to build a model capable of predicting whether future loans will be paid or not. After a long process of data wrangling and analysis of variables, we built and evaluated three possible models, two logistic regressions and a random forest. The final conclusion is that the applied models did not have a satisfactory result and further studies are needed in order to develop something that is applicable in the real world.
Several reasons can be raised as to why the model did not perform properly and with that we were able to define the path for future studies. First, given the way in which the data on fully paid and charged off loans are mixed, I can conclude that a larger set of observations would be necessary in order to properly train the model. In addition, for the same reason, more complex and computationally more demanding models such as deep neural networks and ensemble models can be applied, with the expectation of obtaining better results. A third point is that the losses of each incorrectly detected fully paid loan are not equal to the losses of each incorrectly detected charged off loan and, therefore, the weights given to these errors in the final assessment metric should not be the same. If we had information on the difference between potential losses, we could use the F1 score with an appropriate beta as a measure of accuracy and that would certainly lead to better models.