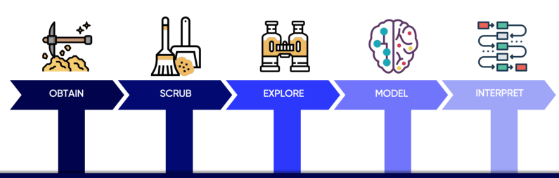
***Credit Score Analysis***

***For***

***Credit One***



[***Zumel and Mount Framework*** 3](#_Toc34865600)

[**Goal** 3](#_Toc34865601)

[**Collect and manage data** 4](#_Toc34865602)

[**Build the model** 5](#_Toc34865603)

[**Evaluate and critique the model** 5](#_Toc34865604)

[**Present results and document** 6](#_Toc34865605)

[**Deploy and maintain the model** 6](#_Toc34865606)

[***BADIR Framework*** 7](#_Toc34865607)

[**Business question** 7](#_Toc34865608)

[**Analysis plan** 7](#_Toc34865609)

[**Project Plan** 7](#_Toc34865610)

[***Flow Chart*** 10](#_Toc34865611)

# ***Zumel and Mount Framework***

## **Goal**

Improve the credit score calculation to determine if potential customers will pay on time or will default the payments of the loan, by analyzing different attributes associated with the consumer such as gender, education, marital status, age, and payment history within next classification:



***The better your score, the cheaper it is to obtain a loan***

## **Collect and manage data**

|  |  |
| --- | --- |
| **Attribute** | **Value** |
| Default Payment | 1 = Yes  0 = No |
| Given Credit Amount | Dollar amount for consumer credit |
| Gender | 1 = male 2 = female |
| Education | 1 = graduate school 2 = university  3 = high school 0, 4, 5, 6 = others |
| Marital Status | 1 = married 2 = single  3 = divorce  0=others |
| Age | Consumer age in years |
| History of Past Payment | 6 months payment records from 09/2005 - 04/2005 -2: No consumption -1: Paid in full 0: The use of revolving credit 1 = payment delay for one month 2 = payment delay for two months …  8 = payment delay foreight months  9 = payment delay for nine months and above |
| Bill Statement Amount | Amount of bill statement from 09/2005 - 04/2005 |
| Previous Payment Amount | Amount of previous payment from 09/2005 - 04/2005 |
| Level of Debt | Calculation of Debt Rate based on monthly incomes against current line of credits. If you are using a lot of your available credit, this may indicate that you are overextended |

## **Build the model**

**Linear Regression**

* What will be the customer payment for next month?
* Which attributes (given credit amount, gender, education, marital status, age, history of past payment, bill statement amount, previous payment, level of debt) are more important in deciding if the customer will repay or not next month loan?
* How the history of past payments are correlated with previous payment amount for monthly default?

**Classification**

*Logistic Regression*

* How does the probability of default change based on customer gender and age?
* How does the probability of default change based on customer education and history of past payment?
* Do the age and marital status factors affect the credit score?

## **Evaluate and critique the model**

The assessment will focus on evaluating the overall fit of the model, the significance of each predictor, and the relationship between the default payment and each attribute: given credit amount, gender, education, marital status, age, history of past payment, bill statement amount, previous payment and level of debt.

|  |  |
| --- | --- |
| **Evaluation** | **Description** |
| Type | What type of data to expect (regression/classification) |
| Number of unique values | Variations in the population |
| Number of missing values | How well the data have been collected |
| Mean, median, and mode | What the average client looks like |
| Histograms | How the population is distributed |
| 5 highest and lowest values | What the outliers are or which values are likely to be errors |

\*Outliers are extreme values that don’t reflect the reality in the data. It is usually good practice to remove extreme outliers from a data set.

The data would be separated randomly into two independent sets, called the training and test datasets. The training is the data that we will use to build our model. Test, is the data that we will use to validate the model.

The objective is to validate that the predictive power of the model in the testing sample is close to that in the developing sample. If this is not the case, the variables that present the biggest difference should be evaluated and adjusted: reduce the number of categories, eliminate outliers, or eliminate variables with a high proportion of missing values.

Then to evaluate the performance a Confusion Matrix will be used to show the consequences for the business:

* *False Negative:* Predicting a bad loan when it was a good one. The impact would be a potential loss of profits.
* *False Positive:* Predicting a good loan when it was a bad one. The consequence could be a potential loss of interests and principal and even the addition of recovery costs.

## **Present results and document**

One of the most appealing features in logistic regression is the transparency in the interpretation of the model. It provides the change in the probability of default if the variable changes by one unit. This can help to understand how the probability of default changes if (for example) the savings capital of a client changes from $1,000 to $100,000 EUR. Similarly for all other variables.

Armed with this result we can provide further guidance to the business by giving the impact of each individual explanatory variable to the client’s forecasted probability of default.

## **Deploy and maintain the model**

The best model will be incorporated permanently. However, challenger models should also be created to see if they will work better than the current model. If they do, they can be adopted or some of their details used to amend the current one.

Besides that the *Data Science team* of experts provides assistance at various levels of the modelling process, from training to design to implementation, to validation, to deployment.

# ***BADIR Framework***

## **Business question**

How do you know if a consumer will be defaulted on loan’s payments?

Credit One wants to build custom scoring models to predict more accurately by using internal company data if someone will be n days late on a loan payment out of taking the loan. In order to better understand the risks and opportunities associated with their particular customers and prospects.

## **Analysis plan**

*Goal*

Improve the credit score calculation to determine if potential customers will pay on time or will default the payments of the loan, by analyzing different attributes associated with the consumer such as gender, education, marital status, age, and payment history.

*Hypotheses*

* Higher scores mean fewer defaults and vice versa.
* Age, gender and marital status are factors that don’t affect the scores.
* The more educated biggest limit balance and less default.
* The higher your score, the better terms of a loan.

*Methodology*

**Classification**

*Logistic Regression*

One of the most common, successful and transparent ways to do the required binary classification to “good” and “bad” is via a logistic function. This is a function that takes as input the client characteristics and outputs the probability of default.

* How does the probability of default change based on customer gender and age?
* How does the probability of default change based on customer education and history of past payment?
* Do the age and marital status factors affect the credit score?

## **Project Plan**

*Default data*

Before the analysis begins it is important to clearly state out what defines a default. This definition lies at the heart of the model. Different choices will have an impact on what the model predicts. Some typical choices for this definition include the cases that the client misses three payments in a row, or, that the sum of missed payments exceeds a certain threshold.

In order for the model to be able to make accurate forecasts it needs to see enough examples of what constitutes a default. For this reason it is important that there is a sufficiently large number of defaults in the data. Typically in practice, data with less than 5% of defaults pose strong modelling challenges.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **# Loans** | **Not Default** | **Default** | **% No Default** | **% Default** |
| 30,000 | 23,364 | 6,636 | 77,88 | 22,12 |

*Data collection*

* Structured data in the form of discrete or categorical variables, finite set of possible values such as gender, education, and marital status.
* Structured data in the form of continuous or numeric variables which are integers or real numbers, for example: credit score, limit balance, bill amount.
* Special case of categorical variables are boolean or binary variables which are yes/no or true/false like default payment next month.

Outliers

* One loan opened by a credit amount of 1,000,000.

*Data Patterns*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Age** | | **# Loans** | **Default** | | **No Default** | **% Default** | **% No Default** |
| 20's | | 9,618 | 2,197 | | 7,421 | 22,84 | 77,16 |
| 30's | | 11,238 | 2,276 | | 8,962 | 20,25 | 79,75 |
| 40's | | 6,464 | 1,586 | | 4,878 | 24,54 | 75,46 |
| 50's | | 2,341 | 582 | | 1,59 | 24,86 | 75,14 |
| 60's | | 314 | 89 | | 225 | 28,34 | 71,66 |
| 70's | | 25 | 7 | | 18 | 28,00 | 72,00 |
|  | |  |  | |  |  |  |
| **Marital Status** | | **# Loans** | **Default** | | **No Default** | **% Default** | **% No Default** |
| Single | | 15,964 | 3,341 | | 12,623 | 20,93 | 79,07 |
| Married | | 13,659 | 3,206 | | 10,453 | 23,47 | 76,53 |
| Divorce | | 323 | 84 | | 239 | 26,01 | 73,99 |
| Others | | 54 | 5 | | 49 | 9,26 | 90,74 |
|  | |  |  | |  |  |  |
| **Gender** | | **# Loans** | **Default** | | **No Default** | **% Default** | **% No Default** |
| Male | | 11,888 | 2,873 | | 9,015 | 24,17 | 75,83 |
| Female | | 18,112 | 3,763 | | 14,349 | 20,78 | 79,22 |
|  | |  |  | |  |  |  |
| **Education** | | **# Loans** | **Default** | | **No Default** | **% Default** | **% No Default** |
| Graduate School | | 10,585 | 2,036 | | 8,549 | 19,23 | 80,77 |
| University | | 14,030 | 1,237 | | 12,793 | 8,82 | 91,18 |
| High School | | 4,917 | 3,330 | | 1,587 | 67,72 | 32,28 |
| Others | | 468 | 33 | | 435 | 7,05 | 92,95 |
| **LIMIT\_BAL** | **EDUCATION** | | |
| 1,000,000 | graduate school | | |
| 800,000 | university | | |
| 800,000 | graduate school | | |
| 780,000 | graduate school | | |
| 780,000 | university | | |
| 760,000 | high school | | |
| 750,000 | university | | |
| 750,000 | high school | | |
| 750,000 | graduate school | | |
| 750,000 | graduate school | | |

|  |  |  |
| --- | --- | --- |
| **Payment Delayed** | | |
| **April** | **September** | **% Delayed** |
| 1,609 | 1,282 | 79,68 |

*Proven / Disproven Hypotheses*

* The older you are, the most difficult to get a loan.
* The older you are, less need to request a credit.
* The youngest you are, the most debts you have.
* The age is not a determinant factor to request a loan.
* No matter the age between 20%-30% of the accounts are in defaulted.
* Marital status is not a factor to default loan, 53,21% of the loans are requested by single people and 45,53% by married people, for a total of 98,74% and around the 20% are in default.
* Most females requested a loan but are 5% under the male default rate.
* The highest loans are requested for the most educated people.
* Less educated people are most in default, 67,72% of high school loans are in default.
* 79,68% of the accounts that were in default in Apr 2005 are in default in Sep 2005.

*Results Confidence*

The predictive model "learns" by utilizing a customer's historical data together with peer group data and other data to predict the probability of that customer displaying a defined behavior in future

Generally speaking, the higher the credit score, the more confident the stakeholder can be of the customer's creditworthiness.

Businesses can specify the factors they want considered in the credit decision process to know almost immediately if they are dealing with a high-risk or low-risk customer.

*Impact on the business*

The greatest benefit of credit scoring is the ability to help make decisions in a fast and efficient way, such as to accept or reject a customer or increase or decrease loan value, interest rate, or term.

The chance that someone with a given score may fall behind can change with shifts in consumer behavior and the economy. For example, people may be more likely to pay a bill late during a recession.

Credit One may periodically revalidate their models to see if they need adjusting or to see if the model’s outputs call for any changes in business strategy.

*Recommendation*

Businesses may continually evolve, so besides the production model Credit One must monitor and back-test the vintage models that were previously used to compare performance. If a new model outperforms the current model, it may become a new champion model.

# ***Flow Chart***

.

