MASTER'S THESIS

Interuniversity Master in Statistics and Operations Research UPC-UB

Title:	**Kiva.org: Exploring Relationships in Online Crowdfunding**
Author:	**Marc Valentí**
Directora:	**Dra. Inés Alegre**
Ponent:	**Dr. Xavier Tort**
Department:	**Estadística I Investigació Operativa**
University:	**Universitat Politècnica de Catalunya** **Universitat de Barcelona**

Academic year: 2017-2018

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Estadística
Departament d'Estadística i Investigació Operativa

Interuniversity Master in Statistics and Operations Research UPC-UB

Kiva.org: Exploring Relationships in Online Crowdfunding

by Marc VALENTÍ

Internet has completely altered communications around the world. The sector of crowdfunding has widely benefited from its appearance. In this thesis, we aim to contribute to the current research in online crowdfunding by working with data from Kiva.org, one of the world's main online lending platforms.

In it, we review the current research to expand the present knowledge of it by focusing on several independent aspects. We focus on (1) how during the decision process of the lender, similarities with the borrower (such as gender, country or occupation) are relevant. Attention is also brought to the (2) descriptions of the loans, aiming to observe how they differenciate among each other and how they evolve over time. Finally, (3) Machine Learning is used to perform Image Recognition in order to encode borrower face emotions from their pictures, while concluding its importance in a rapid funding rate of a loan.

Preface

One year ago, after finishing the first year of this Master, I was given the opportunity to move to Berlin. I could not let escape it, so I decide to postpone this Thesis for a Semester.

I spent time thinking on what I wanted to become the topic of my Thesis. I had some ideas in mind, but I headed to Inés' office to know more about what she was interested in doing. And it all clicked. Being very interested in the field of behavioral economics, I found the topic of crowdfunding very attractive. On the other hand, Kiva had become extremely friendly in allowing third parties using their data, being completed and documented.

From the very beginning I had very clear that I wanted this Thesis to be public and completely reproducible. The repository `https://github.com/mvalenti12/TFM` contains all the code and instructions to reproduce this entire Thesis. [1]

On the other hand, I wanted to folow the best practices and thus have written[2] this thesis in LaTeX, while storing all the bibliography in Mendeley and having a proper integration.

Now that is over, I feel very proud of all the effort behind. What is written in the next pages is just a little summary of the work that has been behind.

Many people that have contributed to this Thesis and I would like to state my most sincere thanks.

First of all, to my family; *Gràcies per haver-me proporcionat una educació.*

Thanks Inés for all the guidance and help during this Thesis. It has been a great team. Thanks Xavier for providing me feedback and acting as a bridge with the University.

I would like to thank the entire MESIO: professors, friends and colleagues for making a great experience out of this.

Finally, to my King colleagues. I have been learning a lot from you and feel honoured to have your support.

I am not forgetting about you, three employees of Wasabi Sushi in V Jámě 1371/8, Prague. Catching the burglar that stole the computer with *so many uncommitted* progress. [3].

[1] For any questions or feedback, please contact marcv94@gmail.com

[2] This great template was used `http://www.sunilpatel.co.uk/thesis-template`

[3] Do not hesitate to ask for this fun story, recovering my PC after signing more than 20 pages of police statements written in Czech at 1am

Contents

List of Figures

List of Tables

Chapter 1

About Kiva and the Dataset

1.1 About Kiva

Kiva is a 501(c)(3) non-profit organization, founded in 2005 and based in San Francisco, with a mission to connect people through lending to alleviate poverty. [1] It is an online platform focused on micro-loans (with a median loan of 500$) that aims to connect borrowers and lenders.

Since its origin, 1419607 loans have been posted, being 1355316 of them funded, reaching an impressive funding rate of 95.47%. During 2017, 224618 loans were created, raising a total amount of 161.92M$.

On the other side, the number of lenders is 2349175; 183734 of them joining on 2017.

There are four main agents in the platform:

- **Borrowers**, users that request loans. They can be charged interest rate fees from the **Partners**.
- **Partners**, mostly a local microfinance institution acting as a bridge between **Borrowers** and **Kiva**. Can charge interest rate, being reflected on **Borrower**'s side.
- **Lenders**, users that have the option of contributing with at least 25$ to the different available projects. No interest rate is received from the loans.
- **Kiva**, being the platform that connects **the three previous agents**, agregates and transfers the capital.

In 1.1, provided by I. Alegre, a schema of how the platform works is displayed.

[1] See Kiva (*Kiva Website | About Kiva*).

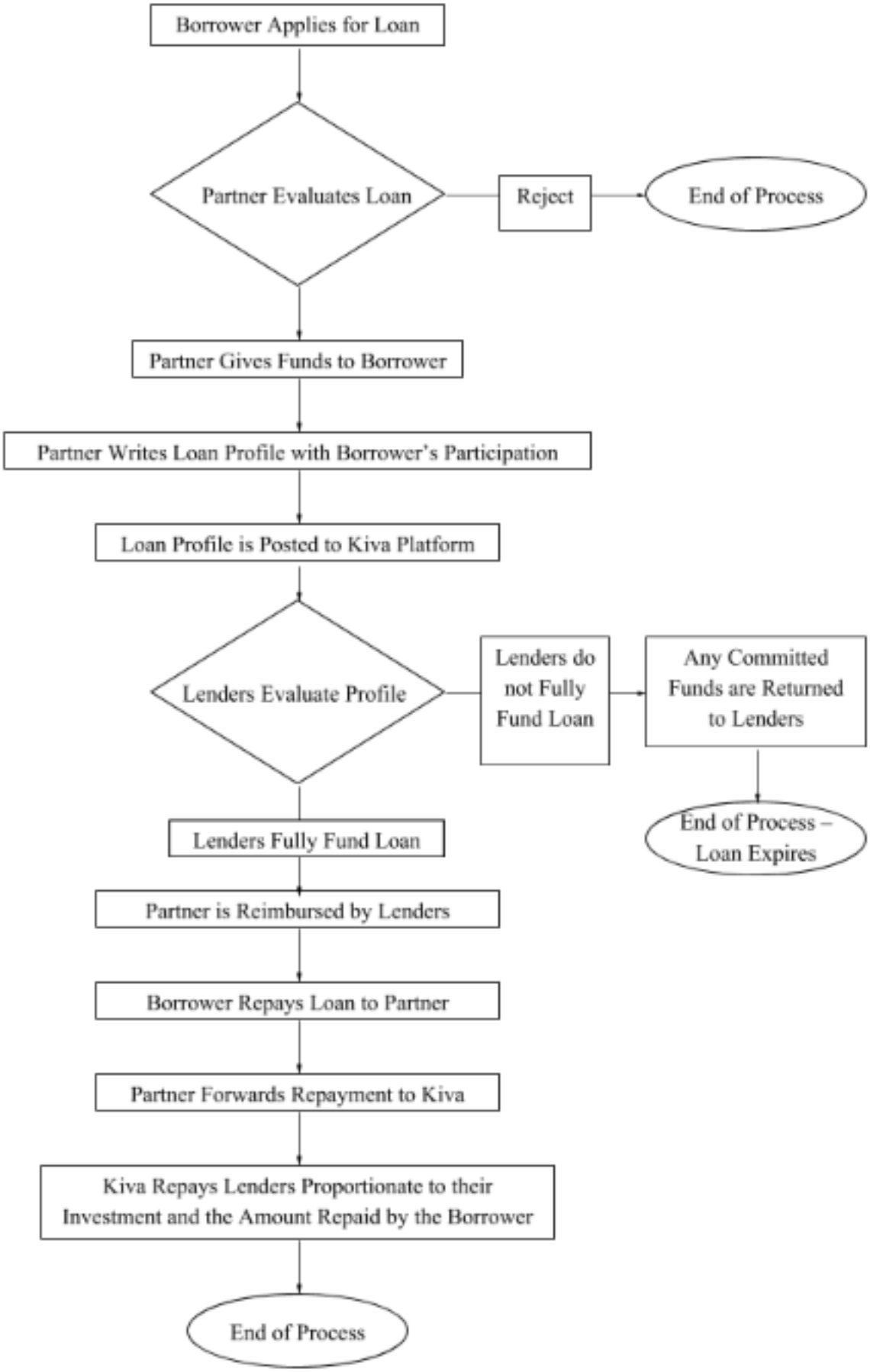


FIGURE 1.1: Kiva Schema

1.2 Accessibility and Description of the data

Kiva aims to offer facilities to developers to access the data. In the following subsections, we present the two approaches followed to extract and use the data presented in this thesis: a RESTful web-service API and using Data Screenshots.

1.2.1 RESTful web-service API

The RESTful web-service API aims at offering real-time data by doing requests. It becomes a great alternative and prevents the tedious web-scrapping.

The documentation that can be found here: http://build.kiva.org/docs/conventions/restful

> REST means that our API is resource-oriented where the method calls are actually URLs that point to data you fetch or modify. The "ful" part of this means that we are not that picky about a formal definition of REST and actually quite happy to use our intuition about things as we go along.
>
> At Kiva, our resources are generally the nouns you think of when visiting our site – Loans, Lenders, Field Partners, and Lending Teams. Thus, in true RESTful fashion, if you wanted to fetch data about a loan at Kiva, you'd make an HTTP GET request accompanied by the proper URI for the loan about which you want information. To create a loan, you might send a POST request to a similar URI. Since the Kiva API currently only supports fetching data, you only need to concern yourself with GET requests for now.
>
> Generally, the RESTful nature of an API is wrapped up on the manner in which that API lets users modify, remove, or fetch data. However, it has some healthy implications for how other very important parts of the API work too.

Regarding Authorization, it uses Oauth 1.0a protocol to control application access to protected resources.[2]

[2] For further information on Oauth check: http://build.kiva.org/docs/conventions/oauth

1.2.2 Data Snapshots

Since doing multiple requests on the Kiva API can be overbearing, this alternative becomes more handy when bigger volumes of data are requested. This is the ideal format for analyses and visuals, and the one used in this thesis. The snapshot presented was taken on the 20th August 2018, containing data from its origin: 16th April 2006 until 31st December 2017.

It contains three different datasets:

- loans.csv Contains information about the loans
- loans_lenders.csv Contains the mapping of all the lenders of a given loan
- lenders.csv Contains information about the lenders.

1.2.2.1 Description of loans.csv

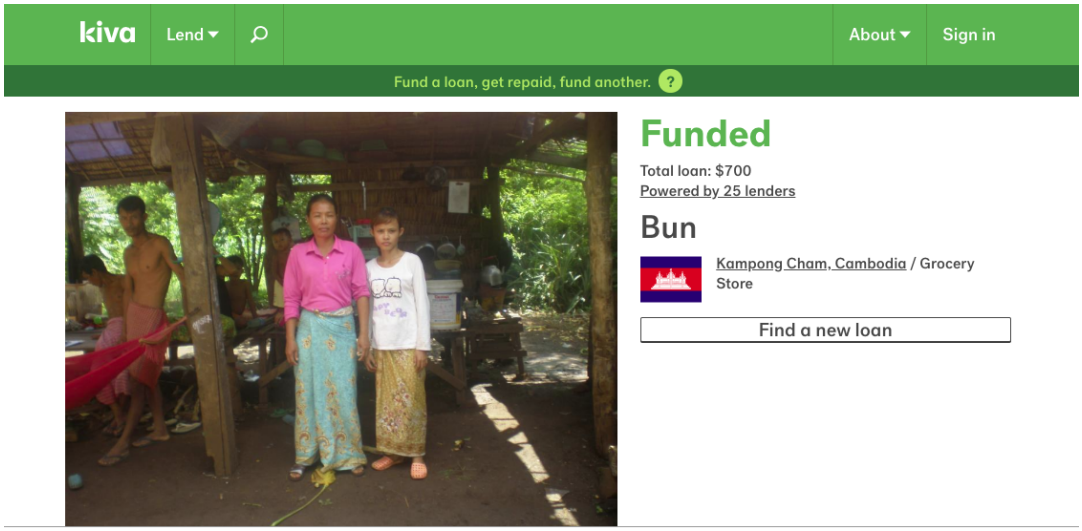
The loans.csv file is the one with highest size: 2.12 GB. It contains 1419607 rows and 34 variables. To provide an example of how does the information in the website appear on the dataset we include, for a given loan (in this case the loan id = 120122, that can be accessed at `https://www.kiva.org/lend/120122`) in *italics*, what is the value encountered. Also, visual support is offered in Figure 1.2, showing a screenshow of the web browser.

More concretely, the variables found on every file are:

- **id (1)** *120122*, Unique identifier of every loan.
- **name (2)** *Bun*, Name of the loan requester.
- **original_language (3)** *English*, Original language of the loan. See Table 1.1.
- **original_description (4)** *Bun T., 42, is married and lives in Kampong Cham Province with her six children. Her husband moved to work in Phnom Penh City as a construction worker with an income of US$8 per day. She has made up her mind that she wants to take a loan of US$700 to create a business of selling groceries at her house with the assistance of her children. If she succeeds in her business plan, she will find a job suitable for her husband to work back in his hometown.*, Description listed in the loan.
- **translated_description (5)** *NA* When the **original_description (4)** is not in english, the translation to english appears here.

- **funded_amount (6)** *700*, Amount the loan managed to fund.
- **loan_amount (7)** *700*, Amount required to fund.
- **status (8)** *funded*, The different status are: fundraising, funded, in_repayment, paid, defaulted, refunded.
- **image_id (9)** *347599*, Unique identifier for the image. There are 1242540 identifiers.
- **video_id (10)** *NA*, Unique identifier for the video. There are 495 different identifiers.
- **activity (11)** *Grocery Store*, There are 163 different activities.
- **sector (12)** *Food*, There are 15 different sectors. See Table 1.2.
- **use (13)** *To buy groceries for her new grocery store*, Brief description; what the loan is aimed for.
- **country_code (14)** *KH*, Country code. There are 94 different categories.
- **country_name (15)** *Cambodia*, Country Name. See Table 1.3.
- **town (16)** *Kampong Cham Province*, Town
- **currency_policy (17)** *not shared*
- **currency_exchange_coverage_rate (18)** *NA*
- **currency (19)** *USD*, Currency the Loan is in. There are 75 different factors.
- **partner_id (20)** *9*, Unique identifier for a partner. There are 464 different partners.
- **posted_time (21)** *1246610728*, Time in msts the loan was posted.
- **planned_expiration_time (22)** *NA*, Time (and date) in msts the loan is expected to expire.
- **disbursed_time (23)** *1244703600*, Time (and date) in msts on which the borrower got the loan.
- **funded_time (24)** *1248360184* Time (and date) in msts the loan got funded.
- **term_in_months (25)** *17* Time the loan will remain posted.

- **lender_count (26)** *25* Number of lenders that have funded the loan.
- **journal_entries_count (27)** *1*
- **bulk_journal_entries_count (28)** *1*
- **tags (29)** #*Animals* Labels associated.
- **borrower_names (30)** *Bun* Vector containing the names of the borrowers, separated by commas.
- **borrower_genders (31)** *female* Vector containing the gender of the borrowers, separated by commas.
- **borrower_pictured (32)** *true* Vector containing a boolean whether the borrower appears in the picture, separated by commas.
- **repayment_interval (33)** *monthly*. Either *bullet* (111568), *irregular* (529446), *monthly* (777971) or *weekly* (622).
- **distribution_model (34)** *field_partner* Either *field_partner* (1402817) or *direct* (16790).



A loan of $700 helped to buy groceries for her new grocery store.

Bun's story

Bun T., 42, is married and lives in Kampong Cham Province with her six children. Her husband moved to work in Phnom Penh City as a construction worker with an income of US$8 per day. She has made up her mind that she wants to take a loan of US$700 to create a business of selling groceries at her house with the assistance of her children. If she succeeds in her business plan, she will find a job suitable for her husband to work back in his hometown.

More about this loan

About KREDIT Microfinance Institution:

KREDIT Microfinance Institution Plc., is one of Kiva's most established partners in Southeast Asia. The organization empowers the economically-active poor and small entrepreneurs by providing inclusive financial services. Beyond loans, KREDIT offers low-income clients complementary training in debt management, savings and budgeting. In addition, KREDIT partners with NGOs to offer training in primary health care, agricultural techniques and HIV/AIDS awareness.

For more information on KREDIT, please visit its partner page. If you would like to engage more with KREDIT and its borrowers, you can also join its lending team, KREDIT MFI Cambodia.

Lenders and lending teams

Contributing lenders (25)

Marvin · Vic M, Chico, CA, U... · Marketta & T..., Frankfurt, Ge... · Anne, Oslo, Norway · Dan & Wanda, Aurora, IL, U...

David, Bordentown, ... · Jo & Jack, Anchorage, ... · Family, Westerville, ... · Tri, New Brighto... · Brodie and J..., Grosse Point...

Loan details

Loan length:

17 months

Repayment schedule: **Monthly**
Disbursed date: **June 11, 2009**
Currency exchange loss: **N/A**
Facilitated by Field Partner:
KREDIT Microfinance Institution
Is borrower paying interest? **Yes**
Field Partner risk rating: ★★★★½

Field Partner: KREDIT Microfinance Institution

Why Kiva works with this partner:

Kiva chose KREDIT because it serves a high percentage of women, rural and low income borrowers. Additionally, we were impressed by KREDIT's recent addition of a designated team that reaches out to vulnerable groups like families affected by HIV/AIDS.

Time on Kiva: **148 months**
Kiva borrowers: **23,648**
Total loans: **$15,400,225**
Average cost to borrower: **23% PY**
Profitability (return on assets): **1.4%**
Average loan size (% of per capita income): **93.60%**
Delinquency rate: **1.66%**
Loans at risk rate: **2.49%**
Default rate: **0.09%**
Currency exchange loss rate: **0.00%**

More about KREDIT Microfinance Institution
What is a Field Partner?

FIGURE 1.2: Example of a loan on the browser. Loan_id = 120122.

1.2.2.2 Tables of loans.csv

In this subsubsection, a summary of some relevant variables can be found. They group the number of loans (being a total of 1419607), on different variables.

Language	Absolute Value	Frequency (%)	Cumulative Frequency
English	929567	65.5%	65.5%
Spanish	333810	23.5%	89%
French	66468	4.7%	93.7%
Russian	36155	2.5%	96.2%

TABLE 1.1: Number of Loans by Language of Description

Sector	Absolute Value	Frequency (%)	Cumulative Frequency
Agriculture	345950	24.4%	24.4%
Food	322425	22.7%	47.1%
Retail	286887	20.2%	67.3%
Services	103229	7.3%	74.6%
Clothing	81255	5.7%	80.3%
Housing	60109	4.2%	84.5%
Personal Use	51460	3.6%	88.1%
Education	45647	3.2%	91.3%
Transportation	39075	2.8%	94.1%
Arts	28003	2%	96.1%
Construction	18884	1.3%	97.4%
Health	16581	1.2%	98.6%
Manufacturing	15880	1.1%	99.7%
Entertainment	2112	0.1%	99.8%
Wholesale	2110	0.1%	99.9%

TABLE 1.2: Number of Loans by Sector

Country	Absolute Value	Frequency (%)	Cumulative Frequency
Philippines	285336	20.1%	20.1%
Kenya	143699	10.1%	30.2%
Peru	86000	6.1%	36.3%
Cambodia	79701	5.6%	41.9%
El Salvador	64037	4.5%	46.4%
Uganda	45882	3.2%	49.6%
Pakistan	45120	3.2%	52.8%
Tajikistan	43942	3.1%	55.9%
Nicaragua	42519	3%	58.9%
Colombia	33675	2.4%	61.3%

TABLE 1.3: Number of Loans by Country

1.2.2.3 Description of loans_lenders.csv

Regarding the loans_lenders.csv, it acts as a mapping in a one-to-many relationship between loans and lenders. Its size its way lower: 322.7 MB; containing 2349175 rows and two variables. Same as for the description of loans.csv, we include in *italics* the value of the loan 120122.

- **loan_id** *120122*, Unique identifier of every loan.
- **lender_id** *vlasta7274, terrinyc, norm3402, stan7653, brodieandjulie7932, david9344, kazuko6587, denise7998, joannc6734, family6872, t6109, reid, reid, nancy4375, marvin2658, tom5854, tri8862, teamsecular, cristina8587, amanda6636, wandadan8972, anne5946, marketta5407*, all the **lender_id** that have participated in the loan, being comma separated.

1.2.2.4 Description of lenders.csv

Regarding the lenders.csv file, it contains 2262676 rows and 14 variables. The file size is 193MB. Same as before, we will show in *italics* the values for a observation as an example. In this case, the user is david9344 (`https://www.kiva.org/lender/david9344`), that participated in the loan mentioned before. The variables are as follows:

- **id (1)** *david9344*, Unique identifier of every lender.
- **name (2)** *David*, Name of the lender.
- **image_id (3)** *1563180*, Unique identifier of the image of every lender.
- **city (4)** *Bordentown*, City of residence of the lender.
- **state (5)** *NJ*, State of residence of the lender.
- **country_code (6)** *US*, Country of residence of the lender. See Table **??**.
- **member_since (7)** *1193259239*, Date of registration, in msts. This corresponds to Oct 24, 2007.
- **personal_url (8)** *NA*, Personal url.
- **occupation (9)** *computer programmer*, Occupation.
- **loan_because (10)** *Jesus spoke many times about "helping the least of these."*, Reason of participation.

- **other_info (11)** *I work on a computer system for the governnment and help to make it run better.*, Additional information.
- **loan_count (12)** *2117*, Number of loans.
- **invited_by (13)** *Michael*, By whom was invited to Kiva. Unfortunately, it does not correspond to **id (1)**, therefore it is not possible to match.
- **invited_count (14)** *2* Number of invitations accepted.

1.2.2.5 Tables of lenders.csv

The table below shows the Number of Lenders by Country. Unfortunately, 1650424 (70.3%) of the Lenders contain a NULL value in this field.

Country	Absolute Value	Frequency (%)	Cumulative Frequency
United States	591612	25.2%	25.2%
Canada	67970	2.9%	28.1%
India	7203	0.3%	28.4%
Brazil	4200	0.2%	28.6%

TABLE 1.4: Number of Lenders by Country

1.2.2.6 Enriching lenders.csv: Gender

Having no data available for lenders' gender, we suggest an approach to classify them. We are dealing with a classification problem: willing to label every lender with the label *male* or *female*. Our approach is, by using a names dictionary (found in `https://raw.githubusercontent.com/hadley/data-baby-names/master/baby-names.csv`), predict every lender's gender based on its first name. Unfortunately, we cannot report on accuracy on this classification problem since we do not have labelled data. The prediction procedure works as follows:

- Left join the lenders data frame with the names data frame on *name*.
  - If there is a common *name*, keep the *gender* result.
  - If there is no common *name*, apply the **predict_gender** function, as described in Listing 1.1.

```
female_names <- unlist(metadata_names[metadata_names$gender=="female",
                                      'name'])
male_names <- unlist(metadata_names[metadata_names$gender=="male",
                                    'name'])

# both female_names and male_names are vectors containing female names and male
    names

predict_gender <- function(x){
  # initiate the counter of male to 0
  male <- 0
  # initiate the counter of female to 0
  female <- 0
  # split the string into different ones.
  # this is done to identify couples that are registered with both a female and
      male name
  # e.g "Jose y Anna"
  splitted <- unlist(str_split(x,"-|\n| "))
  # for every splitted string
  for (i in 1:length(splitted)){
    # if there is a match with any of the male names in the dictionary,
        increase the male counter
    if(splitted[i]%in%male_names){
      male <- male + 1
      # if there is a match with any of the female names
      in the dictionary, increase the female counter
    } else if (splitted[i]%in%female_names){
      female <- female + 1
    }
  }
  # return the result
  return(ifelse(male*female>=1,
                "couple",
                ifelse(male>=1,
                       "male",
                       ifelse(female>=1,
                              "female","NA"))))}
```

LISTING 1.1: R Code to predict a Lender Loan based on First Name

After applying the prediction procedure for gender, the results are summarized in Table 1.5. There is an exact name match in 70.39% of the cases. By applying the **predict_gender** function, the number of gender fields populated reaches 74.81%. As shown in the listing 1.1; by using the **predict_gender** function a new category is created: `couple`. `couple` is returned when in a string there is a match of both a name labelled as male and a match of a name labelled as female. *Juan y Alejandra* would be labelled as `couple`.

The approach to obtain lender's gender has differed from other related work. In Galak, Small, and Stephen (2010), their approach consisted of scrapping the pictures of the lenders and then, by using Amazon Mechanical Turk service, two independent AMT users would code their gender.

	couple	female	male	couple (%)	female (%)	male (%)
exact match	0	878344	784012	0%	52.84%	47.16%
predict gender	30303	32170	41971	29.01%	30.8%	40.19%
TOTAL	30303	910514	825983	1.72%	51.53%	46.75%

TABLE 1.5: Enriching Lenders' Gender: Results

1.2.2.7 Enriching lenders.csv: Occupation

Since the variable Occupation was an open string, meaning that the lender could type its occupation, a lot of variation was originated. We are willing to group those categories by using Regular Expressions.
In the *lenders* dataset, the *occupation* does not contain any value for 78.53%. With the suggested grouping, 13.6% end up with a occupation group assigned and 7.87% ends up being the lenders who did have a string in the *occupation* field but did not match any of the suggested *RegEx Pattern*.
The groups, shown by group size and Regular Expressions matches seeked, are displayed in Table 1.6.

The approach to obtain lender's occupation has differed from other related work. In Galak, Small, and Stephen (2010), their approach consisted of using Amazon Mechanical Turk service, where two independent AMT users would assign a description to one of their 15 occupations categories.

Occupation Group	RegEx Pattern	Pattern Count	Group Count	Group Count (%)
arts	book	1419	11591	0.49%
	musician	2323		
	writer	7849		
business	accountant	4335	72671	3.08%
	administrator	3305		
	analyst	5844		
	banker	2118		
	business	15219		
	ceo	1592		
	consultant	14247		
	director	6539		
	econom	1157		
	finance	2600		
	marketing	7226		
	sales	8489		
education	educat	5934	47706	2.02%
	professor	4885		
	teacher	36887		
entrepreneur	entrepreneur	8837	8837	0.37%
health	dentist	739	20005	0.85%
	doctor	2036		
	medi	4607		
	nurse	7671		
	pharmacist	852		
	physician	2531		
	psychologist	1569		
it	developer	5523	47905	2.03%
	it	39415		
	programmer	2967		
law	attorney	3861	9106	0.39%
	law	5245		
retired	retired	26754	26754	1.13%
student	student	76679	76679	3.25%
Other	Non NULL	185911	185911	7.87%
Null Description	NULL	1854554	1854554	78.53%

TABLE 1.6: (Summarized) Occupations of Lenders

Chapter 2

Literature Review and Scope of the Thesis

2.1 Literature Review

Online crowdfunding and especially Kiva, have received a lot of attention by researchers. In Alegre and Moleskis (2017), literature review around crowdfunding is done, focusing on the key management theories around crowdfunding and the principal factors contributing to success for the different crowdfunding models. Following with literature review, in Moleskis and Alegre (2018) significant research gaps on crowdfunding are identified.

Due to its completeness, data from the Kiva platform has been widely used to do research around it. All the papers followed above include data from this platform in their analyses: In Moleskis and Canela (2016), the platform is introduced as well as a **collection of descriptive statistics** about both loans and lenders separately. Following with the use of Kiva's data, Canela (2017), the actors on the Kiva Platform are defined, putting a special interest in Africa and the **role of the partner**. **Loans descriptions** are examined in Allison et al. (2015), assessing assess how the linguistic cues in microloan entrepreneurial narratives impact funding outcomes.

Exploring the **lender-borrower relationship**, Burtch, Ghose, and Wattal (2013) evaluate the roles of geographic distance and cultural differences have on lenders' decisions about which borrowers to support. The focus is done on language, difference in GDP, physical distance, cultural differences. Following with the **lender-borrower relationship**, Galak, Small, and Stephen (2010) explores three dimensions of social distance (gender, occupation, and first name initial) between the two agents, concluding that lenders prefer to give to those who are more like themselves. In Jenq, Pan, and Theseira (2011), **objective loan information, pictures and textual descriptions** are investigated as determinants of individual charitable giving. In it, Research assistants

were asked to assess each photograph for qualities such as the borrower's appearance, age, gender, perceived honesty, and skin color on.

Not using Kiva's dataset, Pope and Sydnor (2011) use data from the website Prosper.com, a leader in online peer-to-peer lending in the United States to end up giving evidence of **significant racial disparities** by coding variables from pictures and text descriptions. Following with Propsper.com dataset, Lin and Viswanathan (2013) explore on the **"home bias" topic**, the tendency that transactions are more likely to occur between parties in the same geographical area, rather than outside.

Neither using data from Kiva nor Prosper.com but from Kickstarter, Greenberg and Mollick (2015) put an emphasis on the figure of the female founder. On the lines of the **lender-borrower relationship**, they realize that "(female backers) tended to back projects founded by female founders, and this was especially true in technology".

Not in the academic area but worth mentioning, there was recently a **Kaggle Kernel**[1] based on the Kiva dataset.

[1] https://www.kaggle.com/kiva/data-science-for-good-kiva-crowdfunding/version/4

2.2 Scope of the thesis

The main objective of the thesis is to **contribute to the current research in online crowdfunding** by providing insights and analysis on Kiva dataset. By considering Moleskis and Alegre (2018) and for the authors will, the following hypothesis are to be validated:

On the relationship between lenders and borrowers:

H1: There are other factors than the ones mentioned in previous research that contribute in the lender decision process.

On descriptions:

H2: Every partner has a template for their descriptions; being descriptions distinguishable across partners.

H3: Partners may change their description template over time and copy other partners.

On borrowers' image:

H4: Machine Learning can be used to extract expression labels on images.

H5: The borrowers face expression on the image has an impact on the loan performance.

Last but not least, the main focus and value of this Thesis is the right (both technically and when applicable) application of the different methodologies.

Chapter 3

The lender-borrower relationship; who lends to whom?

3.1 Introduction

This chapter is focused on the first hypothesis:

> **H1:** There are other factors than the ones mentioned in previous research that contribute in the lender decision process.

Inside this hypothesis, several sub-hypothesis were established:

- **H1.1**: Gender: Is there any relationship between the gender of the borrowers and the gender of the lenders? How strong is it?
- **H1.2**: Geographical area: Is there any relationship between the nationalities of the borrowers and the nationalities of the lenders? Is is stronger for any case?
- **H1.3**: Occupations: Is there any relationship between the sector of the borrowers and the occupations of the lenders? Is is stronger for any case?

The current research on the **lender-borrower relationship** can be summarized in the following bullets:

- Using data from Kiva, three dimensions of social distance **(gender, occupation, and first name initial)** are analyzed to conclude that lenders prefer to give to those who are more like themselves Galak, Small, and Stephen (2010).
- Using data from Prosper.com, the importance of geographical area is studied (with the **"home bias" topic**), to conclude that transactions are

more likely to occur between parties in the same geographical area, rather than outside Lin and Viswanathan (2013).

- Using data from Kickstarter, they realize that projects founded by female founders are tended to be backed by more females than projects not founded by females Greenberg and Mollick (2015).

It will contain mostly exploratory analysis, as the conclusions we would get would not be more different to Galak, Small, and Stephen (2010).

3.2 Data Collection

The procedure for extracting data turned out to be computationally costly and not easily scalable. In order to obtain, for a given loan, information about its lenders, the function, named `obtain_lenders_data` was created, working as follows:

- The variables of interest from the lenders are transformed to **one-hot encoding** by using the library `caret`. They are saved as a `data.table` object called *aux_lenders*; assigning `lender_id` as `key`.
- We obtained in the dataframe with source loans_lenders.csv a string of comma separated values with the unique identifiers of lenders. By using the function `stringr::str_split`, we could separate the lenders. For eficiency reasons, we saved all the lenders in a `data.table` object named *lenders_info_expanded_hot*; assigning `lender_id` as `key`.
- Getting rid of the efficiency of `data.table` objects, *lenders_info_expanded_hot* is left-joined to *aux_lenders*. By adding all the colums (that are **one-hot encoded**), each column sum then obtains the number of lenders that had that feature.

.

Now that the function itself was optimized, we seeked for different alternatives to retrieve the information for all the loans. The more efficient was to parallellize it was using the library `parallell`. The procedure worked as follows (see R code in Listing 1.1):

- The number of CPU cores on the current host is identified by using the `detectCores` function.

- A set of copies of R running in parallel and communicating over sockets are created using the `makeCluster` function.
- All the needed variables are exported in the global environment (aka 'workspace') of each node by using the `clusterExport` function.
- The custom function `obtain_lenders_data` was applied in parallell by making use of the `parSapply` function.
- The workers are shut down by calling `stopCluster`. It is good practice to do this.

```
1
2  obtain_lenders_data<-function(lenders_vector){
3    aux_lenders<-(unlist(str_split(gsub(' ','',lenders_vector),',','))) %>%
4      as.data.table() %>%
5      set_colnames(c("id"))
6    aux_lenders[lenders_info_expanded_hot,on="id",nomatch=0] %>%
7      select(-id) %>%
8      colSums(na.rm=TRUE) %>%
9      as.vector() %>%
10     return()
11 }
12
13 no_cores <- detectCores()
14
15 clust <- makeCluster(no_cores)
16
17 clusterExport(clust, "obtain_lenders_data")
18 clusterExport(clust, "%>%")
19 clusterExport(clust, "str_split")
20 clusterExport(clust, "as.data.table")
21 clusterExport(clust, "set_colnames")
22 clusterExport(clust, "lenders_info_expanded_hot")
23 clusterExport(clust, "select")
24
25 dt_complete <- t(parSapply(clust,
26   # might be reccommended to use subsets of lenders
27   # in my local machine, this subset of 50k observations would take 8h
28                              loans_lenders$lenders[200001:250001],
29                              obtain_lenders_data,
30                              USE.NAMES = FALSE))
31
32 stopCluster(clust)
```

LISTING 3.1: R Code to extract lenders information from a loan

We end up collecting information for a sample of 250000 loans. Unfortunately, we were willing to collect the most recent ones, but a bug in the code did not sort the results. Due to the collection time (250000 loans were more than 40h in the local machine) we considered replicating the results for a further iteration but not for this Thesis.

3.3 Methodology

Based on the nature of the data, different approaches will be suggested.

In the case of **gender**, both variables are continuous with a range [0,1]. They are the result of a ratio of the females over females and males. It could be interpreted as failures and successes, accounting thus for the weights (as the number of attempts differs within individuals). Different **regressions** are suggested.

Regarding **occupation** and **countries**, both categories contain categorical variables. This will involve be working with **Contingency tables**, exploring two alternatives to their visualization: **Correspondence Analysis** or **Mosaic Plots**.

3.3.1 Continuous Data with range [0,1]: Gender

Since the number of borrowers added complexity to the estimation of this subsection, we only consider the loans with a single borrower. From now on, to refer to gender as a variable we use the proportion of female individuals out of the whole sample. Regarding the data,

- Regarding **borrowers**, there are way more loans only with female borrowers (174854) than male borrowers (54660).
- Regarding **lenders**, the values are more distributed because the denominator is a higher number. The mode is 0.5, being left skewed. When looking at the extremes; that is loans funded fully by males or females, there are way less loans funded only by female lenders (8623) rather than loans funded only by male lenders (22068).

Even though different models were specified (mostly trying to reproduce Greenberg and Mollick (2015) work explaining how female borrowers were mostly funded by female lenders (and this would differ based on the sector), but none of the models could not be validated. Since results in this area were already shown in the previous paper, we decided to focus on other areas instead. [1].

[1]The Github repository contains many work done on this area so it can be reused in the future

3.3.2 Categorical Data: Countries and Occupations

In the following paragraphs, the methodology used in the current and next subsection is introduced. Both **Correspondence Analysis (CA)** and **Mosaic Plots** are used in doing exploratory data analysis on occupation and country for the Lender-Borrower Relationship.

Regarding **Correspondence Analysis**, its main goal is to display or summarise a set of data in a two-dimensional graphical form. Nagpaul (1999) provides a very good definition of the method:

> **CA** transforms a data table into two sets of new variables called factor score (obtained as linear combinations of, respectively the rows and columns): One set for the rows and one set for the columns. These factor scores give the best representation of the similarity structure of, respectively, the rows and the columns of the table. In addition, the factors scores can be plotted as maps, that optimally display the information in the original table. In these maps, rows and columns are represented as points whose coordinates are the factor scores and where the dimensions are also called factors, components (by analogy with pca), or simply dimensions. Interestingly, the factor scores of the rows and the columns have the same variance and, therefore, rows and columns can be conveniently represented in one single map. In correspondence analysis, the total variance (often called inertia) of the factor scores is proportional to the independence chi-square statistic of this table and therefore the factor scores in **CA** decompose this χ^2 into orthogonal components.

On the other hand, we follow the approach suggested by Zeileis, Meyer, and Hornik (2007) to provide a simple but powerful visualization on **Mosaic Plots**. What is coloured in the plots are the Pearson Residuals.

> To fix notations, we consider a 2-way contingency table with cell frequencies $[n_{ij}]$ for $i = 1, ..., I$ and j = $1, ..., J$ and row and column sums $n_{i+} = \sum_{i+} n_{ij}$ and $n_{+j} = \sum_{+j} n_{ij}$ respectively. Given an underlying distribution with theoretical cell probabilities π_{ij} , the null hypothesis of independence of the two categorical variables can be formulated as
>
> $$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}. \tag{3.1}$$

The estimated expected cell frequencies under H_0 are $\hat{n}_{ij} = n_{i+}n_{+j}/n_{++}$. As well-established in the stastical literature, a very closely related hypothesis is that of homogeneity which in particular leads to the same expected cell frequencies and is hence not discussed explicitely below. The probably best known and most used measure of discrepancy between observed and expected values are the Pearson residuals:

$$r_{ij} = \frac{n_{ij} - \hat{n}_{ij}}{\sqrt{\hat{n}_{ij}}}. \tag{3.2}$$

The most convenient way to aggregate the I x J residuals to one test statistic is their sum of squares

$$\chi^2 = \sum r_{ij}^2 \tag{3.3}$$

because this is known to have an unconditional limiting χ^2 distribution with (I x 1)(J x 1) degrees of freedom under the null hypothesis.

In the Mosaic Diagram, the Lenders variables are shown on the horizontal axis, whereas the borrowers are shown on the vertical one. All the missing values have been discarded, therefore only the valid fields are displayed. The width of the columns indicate the share the borrower variables represent out of the total. The same happens with the height of the rows: they indicate the share the lender variables represent out of the total. Being the **Standardized Residuals** already explained, one of the most clear is the pair United States (Lenders) - United States (Borrowers). United States seems to represent 40% of the funding of the displayed countries. However, there is a big outlier. United States loans by mostly (more than 50%) Americans. On the other hand, loans from Rwanda receive very few participation of United States (US) lenders, being compensated by a big residual from Great Britain.

In the Correspondence Analysis Biplot, Borrowers are plotted in red, whereas Lenders are plotted in Blue. The proximity of two variables of the same group (either Lenders or Borrowers) indicate similarities in their behaviour. Regarding Lenders (in Blue), closer distance means similar row-profile. Borrowers(in Red) close to each other stands for similar column-profile.

3.4 Conclusions

This Thesis started with this Chapter, but we later found out that there was very few room for innovation in this field. The **Lender-Borrower Relationship** has been one of the most explored fields of Academic Research. Therefore, we decided to focus on other areas, the next sections, that are very innovative and the impact would be way higher. However, the summary of this Chapter 3 would be:

Regarding Gender, the suggested models provided very low explainability. As found in Greenberg and Mollick (2015), Female borrowers received more attention by female lenders, and this was stronger in certain Sectors. However, as mentioned, we decided to focus our effort in other sections with higher impact.

Regarding Occupation and Country, both χ^2 statistics reject the Null Hypothesis of independence, exhibiting then stronger deviations and therefore associations. This suggests that every Country has lending preferences. Finally, Correspondence Analysis has not been able to show *similar (based on history, region or race for example)* countries with *similar lending* behaviours.

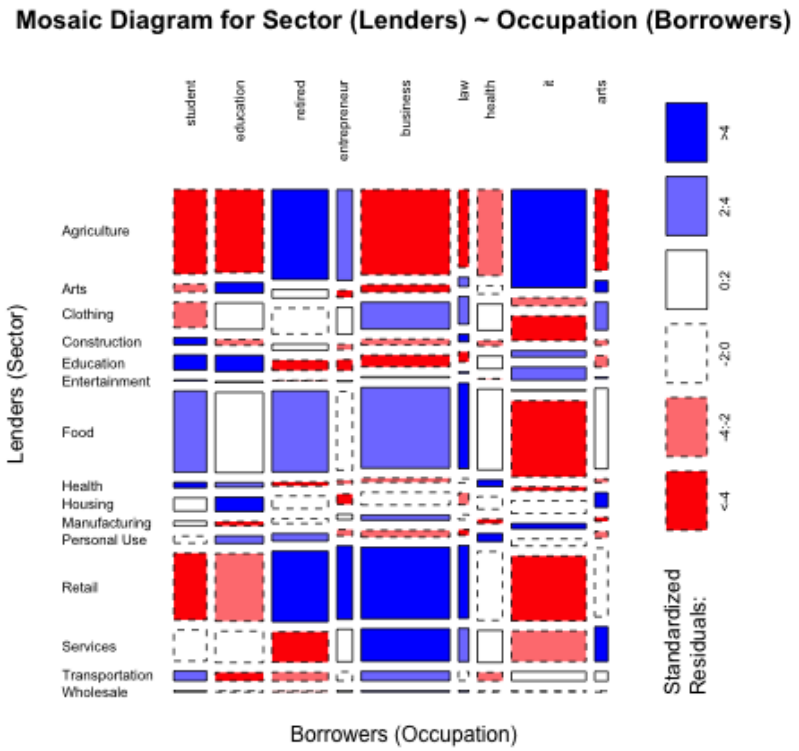


FIGURE 3.1: MCA - Biplot for Occupation (Lenders) - Sector (Loans)

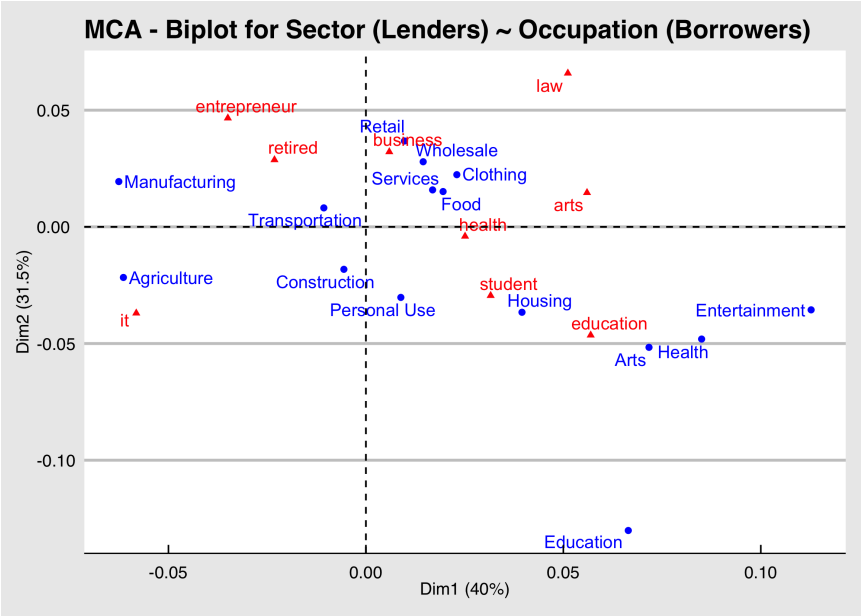


FIGURE 3.2: CA - Biplot for Country (Lenders) - Country (Loans)

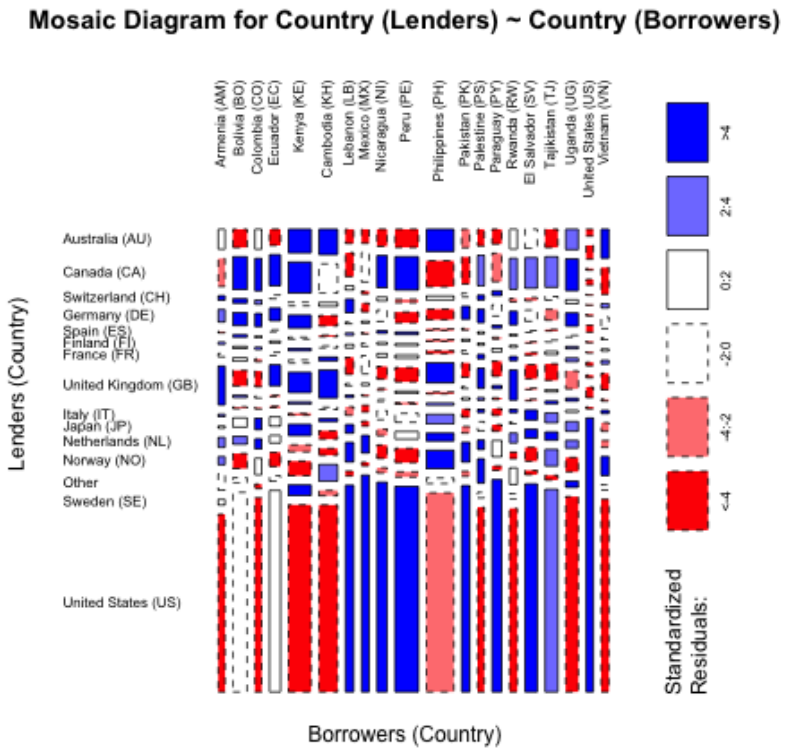


FIGURE 3.3: Mosaic Diagram for Country (Lenders) - Country (Loans)

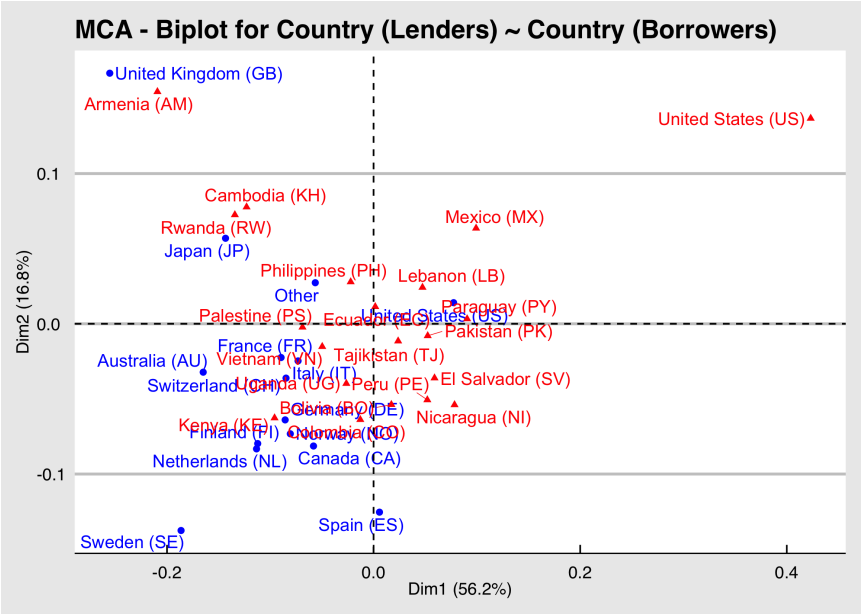


FIGURE 3.4: CA - Biplot for Country (Lenders) - Country (Loans)

Chapter 4

Loans Descriptions

4.1 Introduction

Being the Descriptions of the Loans written by the **partners**, this chapter is focused on working on our description hypothesis:

> **H1:** Every partner has a template for their descriptions; being descriptions distinguishable across partners.

> **H2:** Partners may change their description template over time and copy other partners.

Even though it may sound intuitive that description has great importance on a loan, research around it has been quite scarce. Only (Pope and Sydnor, 2011) focus on descriptions; in this case from Prosper.com.

They have two different approaches: For the short description (one line), they employ two independent Research Assistants per description, they are willing to encode the purpose of a loan. For the longer description, they use a simple text-analysis program, finding correlations between the description (average word-length and average sentence-length, and the percent of words that are misspelled) and picture characteristics. What they end up extracting from the descriptions is then, the log number of total characters, a readability index (which uses word and sentence length), and the percent of words which are misspelled.

Our approach will be completely different. We will be focusing on word frequency; willing to provide a framework to describe the evolution of descriptions over time.

4.2 Text-Processing

As shown in 1.2.2.1, each loan consists of among others, a description. This descriptions is mostly 2 or 3 paragraphs, introducing the borrower and describing the goal of the loan. For the following sections, a subset of data from the country Philippines will be presented. Philippines was chosen since it is the country that has received more loans. However, the methodologies are reproduced for Perú and Kenya two (the second and third countries regarding loan destination). An example of a description could be the one from the loan 120122.

> Bun T., 42, is married and lives in Kampong Cham Province with her six children. Her husband moved to work in Phnom Penh City as a construction worker with an income of US$8 per day. She has made up her mind that she wants to take a loan of US$700 to create a business of selling groceries at her house with the assistance of her children. If she succeeds in her business plan, she will find a job suitable for her husband to work back in his hometown.

Our goal is now to convert this long string into numerical data. Once we have gone through the **Data Acquisition** process, the next step is **Text pre-processing**. In **Text pre-processing**, we are willing to convert our raw descriptions into a matrix.

There are several libraries in R that provide great support in doing this step. This Thesis has used `tm`, but in the meantime `tidytext` was released and `caret` started supporting this format.

In this step, we are willing to remove all the non trivial words and characters of the strings. By applying the **Text pre-processing**, the previous description would result as:

> bun t married lives kampong cham province six children husband moved work phnom penh city construction worker income us per day made mind wants take loan us create business selling groceries house assistance children succeeds business plan will find job suitable husband work back hometown

The next step is the creation of an object of class `Term-Document Matrix` or `Corpus`. A Term-Document Matrix will have nxp dimensions, being n the number of descriptions and p the number of words we decide to retrieve.

X_{nxp} will contain the appearances of word p in the description n. Table 4.1 is an example of a `Term-Document Matrix`, including the previous description.

	business	children	city	groceries	house	husband	income	lives	loan	made	married	province	selling	wants	will	work	able	ahead	continues	grateful
120122	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	2	0	0	0	0
661165	2	0	1	0	0	0	1	1	0	0	0	0	0	0	1	0	1	1	1	1
251336	1	0	0	0	2	0	0	0	2	1	1	0	0	0	1	0	0	0	0	0
423290	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
421136	1	1	0	0	1	0	0	0	1	0	1	0	0	0	1	0	1	1	0	0

TABLE 4.1: A Document-Term Matrix containing Loan 120122

4.3 Visualizing Descriptions

Dealing with high dimensional spaces, we are willing to reduce the dimensionality so that we can gather insights by visualizing a fewer number of dimensions.

After that, a methodology will be shown to describe the evaluation of descriptions over time.

4.3.1 Dimensionality Reduction

Three different approaches will be presented in the following items:

- **Multidimensional Scaling using the Jaccard similarity**: When the decision is to use the Jaccard similarity and to project it to a low-dimensional space by using Metric or Non-Metric Multidimensional Scaling. In this thesis, the `vegan::vegdist` function is used in continuous data to obtain the Jaccard dissimilarity matrix. After that, both Metric Multidimensional Scaling (by using `MASS::cmdscale`) and Non-Metric Multidimensional Scaling (by using `MASS::isoMDS` are applied. Multi Dimensional Scaling aims to visualize the similarities of a Dissimilarity Matrix, being a form of non-linear dimensionality reduction. Very used in the field of Natural Language Processing.

- **Multidimensional Scaling using the Cosine similarity**: When the decision is to use the Cosine similarity and to project it to a low-dimensional space by using Metric or Non-Metric Multidimensional Scaling. The

`lsa::cosine` function is used; using the same procedure to do perform the Multidimensional Scaling. Computing the Jaccard similarity is less computational costly than the Cosine similarity. Very used in the field of Natural Language Processing.

- **Multidimensional Scaling using Euclidean distance (Principal Component Analysis)**: When willing to work with the Euclidean Distance. Two procedures would get the same results: Performing Principal Component Analysis and performing Metric Multidimensional Scaling with the Euclidean Distance. However, Principal Component Analysis uses an orthogonal transformation of the different variables to output principal components, a set of values of linearly uncorrelated variables. The Principal Components are sorted in decreasing variance, while being constrained to be orthogonal to the preceding components. It is less computationally costly than the previous methods. Not much used in the field of Natural Language Processing due to a usual better performance and willingness to work with Jaccard or Cosine similarities.

After doing all the listed procedures, we were aiming to visualize the first two dimensions of a subset of descriptions posted during the month of June of 2016 belonging to Philippines. This is provided in Figure 4.1. In it, the descriptions are plot in a two dimensional space using the techniques previously described. The Non-Metric Dimensional Scaling using the Jaccard dissimilarity and the reduction using Principal Component Analysis show the best aggrupations of clusters.

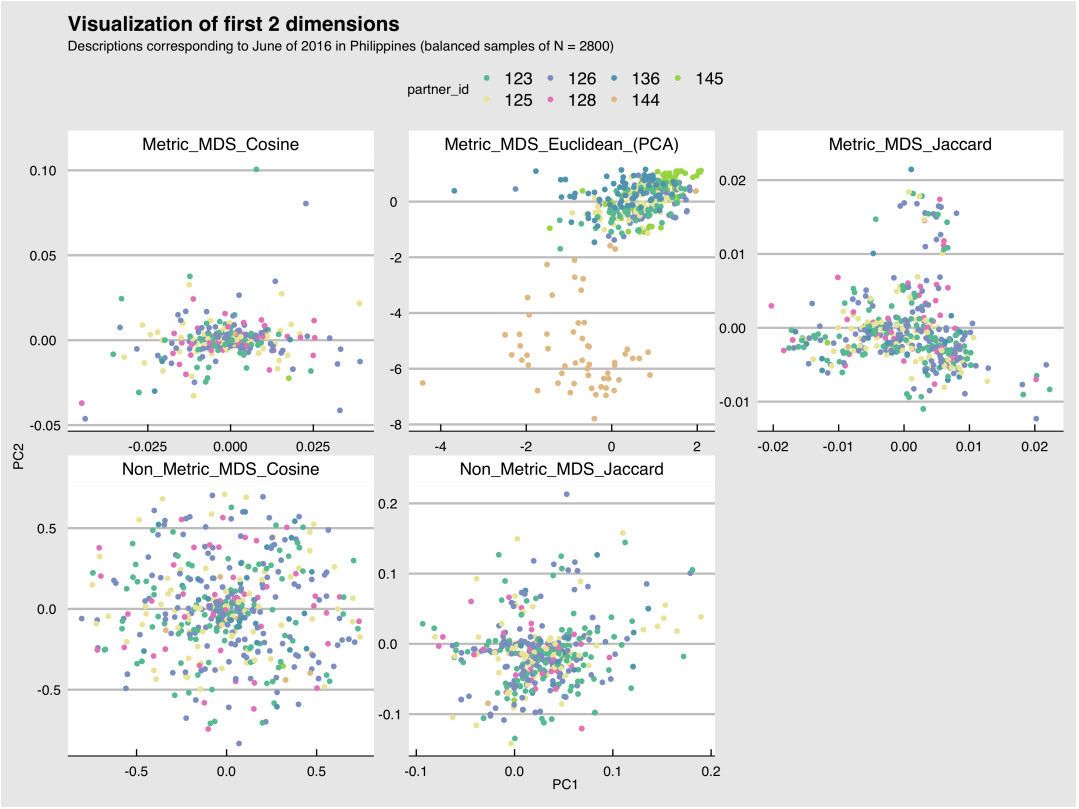


FIGURE 4.1: Descriptions in Philippines, June of 2016. First 2 Dimensions using different distances and methods

Due to the fact that the methodology using Non-Metric Multidimensional Scaling is more computational costly (and not able to do it in the local machine that was used to do this Thesis) and the results are very similar with the methodology with Principal Component Analysis (which has very good scalability), from now on the procedures will be hold with Principal Component Analysis. The `Term-Document Matrix` of Philippines contains more than 285000 documents, being scalability a important topic. Computing a Matrix Distance would require computing more than $4 * 10^9$ pairs of individual distances.

4.3.2 Evaluation of Descriptions Over Time

Having concluded that from this point on due to computationally efficiency, the individual scores from Principal Component Analysis will be used, in this subsection the approach is discussed.

As already mentioned, the Principal Components are sorted in decreasing variance. In Figure 4.2, the percentage of variances explained by each principal component is shown. When choosing the number of dimentions to

retain, the point at which the proportion of variance explained by each subsequent principal component drops off must be chosen. In this case, several options would be possible. Only using the first two principal components would be reasonable, but still has the caveat of only retaining 4.5% of the variance explained. Retaining the first 4 Principal Components is the chosen alternative, since after them the variance drops off. At this point, the explained variance is 7.1%. The last valid alternativewould be retaining the first 6 Principal Components, as the explained variance reaches a plateau after that. It is important to remember that we are using 547 variables in the Principal Component Analysis (all of them scaled and standardized to unit variance), thus the low explained variability.

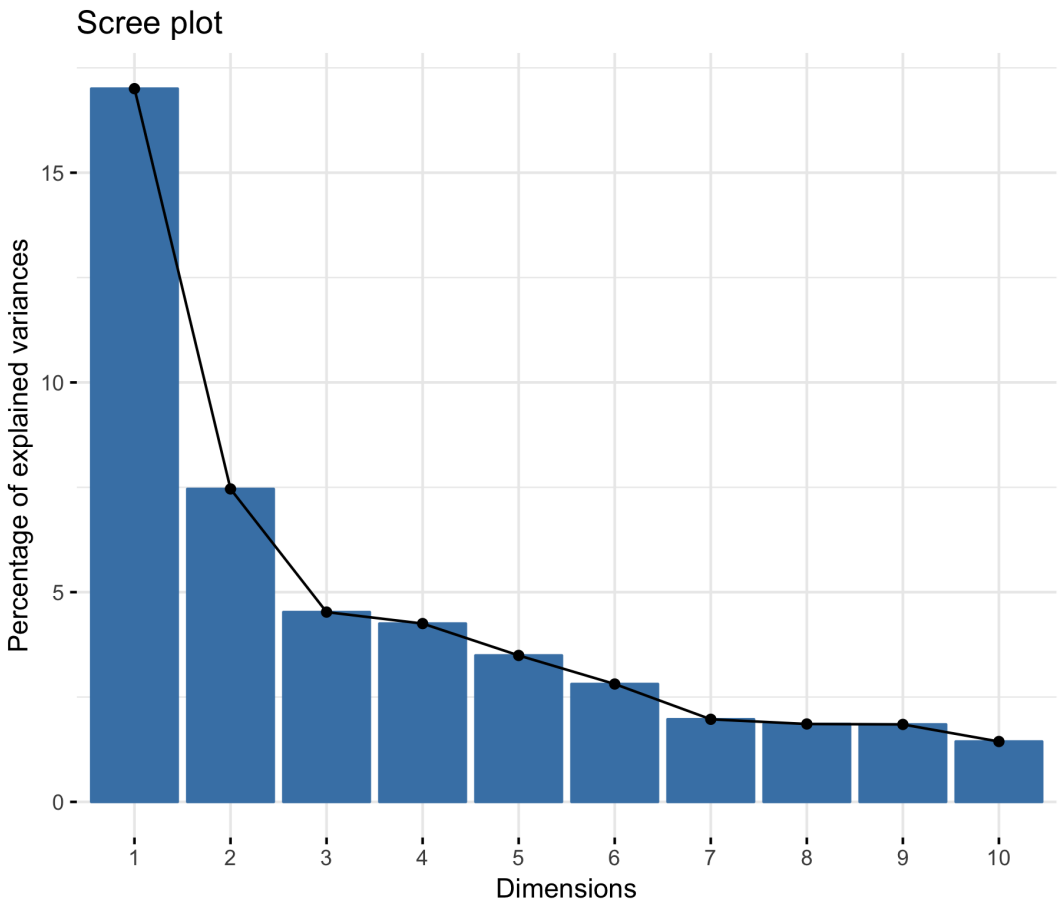


FIGURE 4.2: Scree plot of the eigenvalues

By using the individual scores resulting from the Principal Component Analysis and the month the loan was posted, we provide an evolution over time (in this case, a month is a unit of time) for the descriptions. The individual scores of the descriptions are plotted in the first four dimensional spaces that for Philippines, accounting for 7.1% of the variance. On the other

hand, at every point the average distance (both between and within clusters) is provided for every pair of `partner_ids`. This is summarized in the top-right table of Figure 4.3. On the other hand, the average distance (between and within clusters) is provided for the whole monthly subset, seeing the evolution over time. This is shown on the bottom-right plot of Figure 4.3

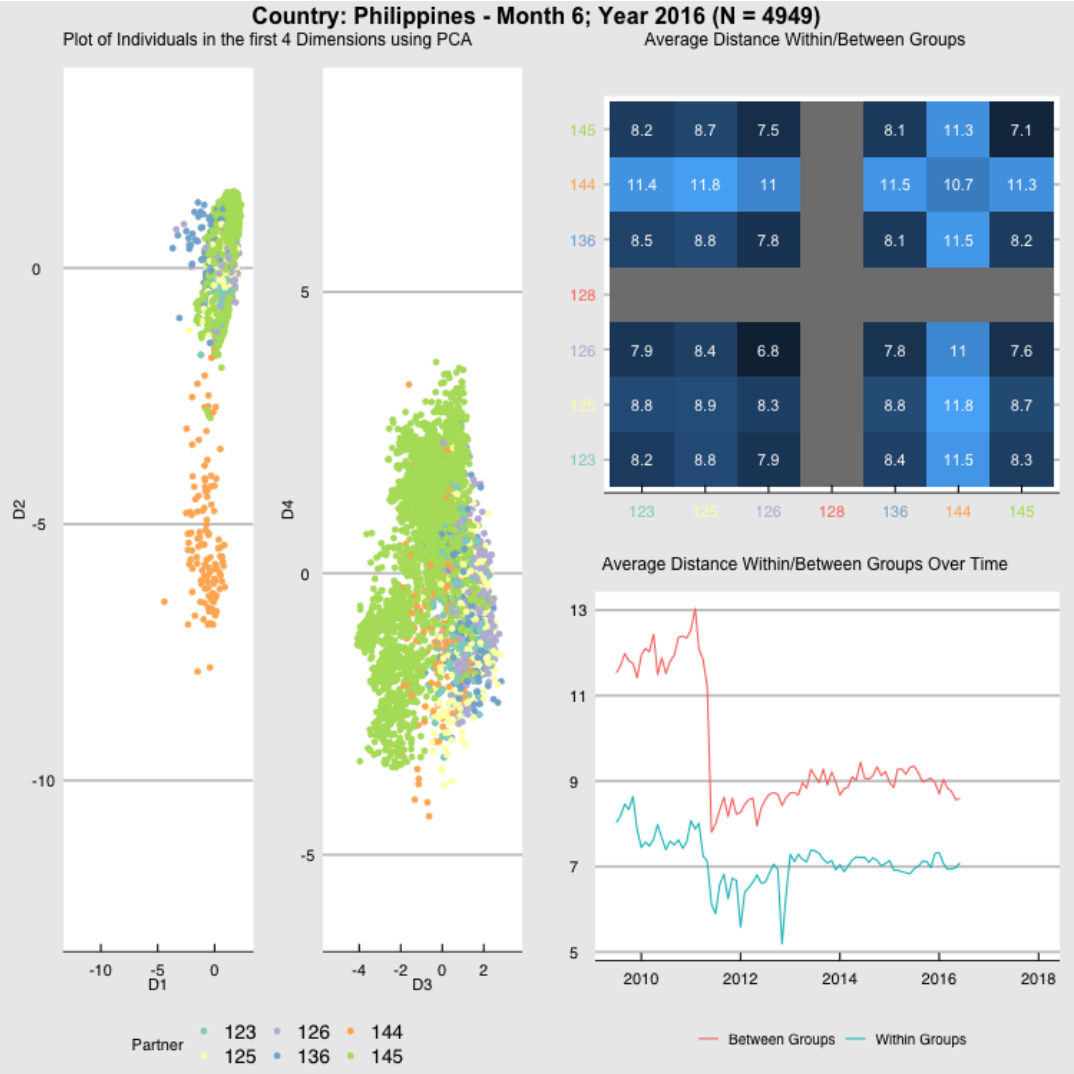


FIGURE 4.3: Descriptions in Philippines, June of 2016. Example Frame

We then provide an animation by using all the different months to see the evolution over time. In Figure 4.4 different frames are shown for the month of July of the years in the period 2012-2017. This subsection is replicated for the three countries that receive more loans: Philippines, Kenya and Perú.[1].

[1]Watching the GIFs available at https://github.com/mvalenti12/TFM/gifs is reccommended

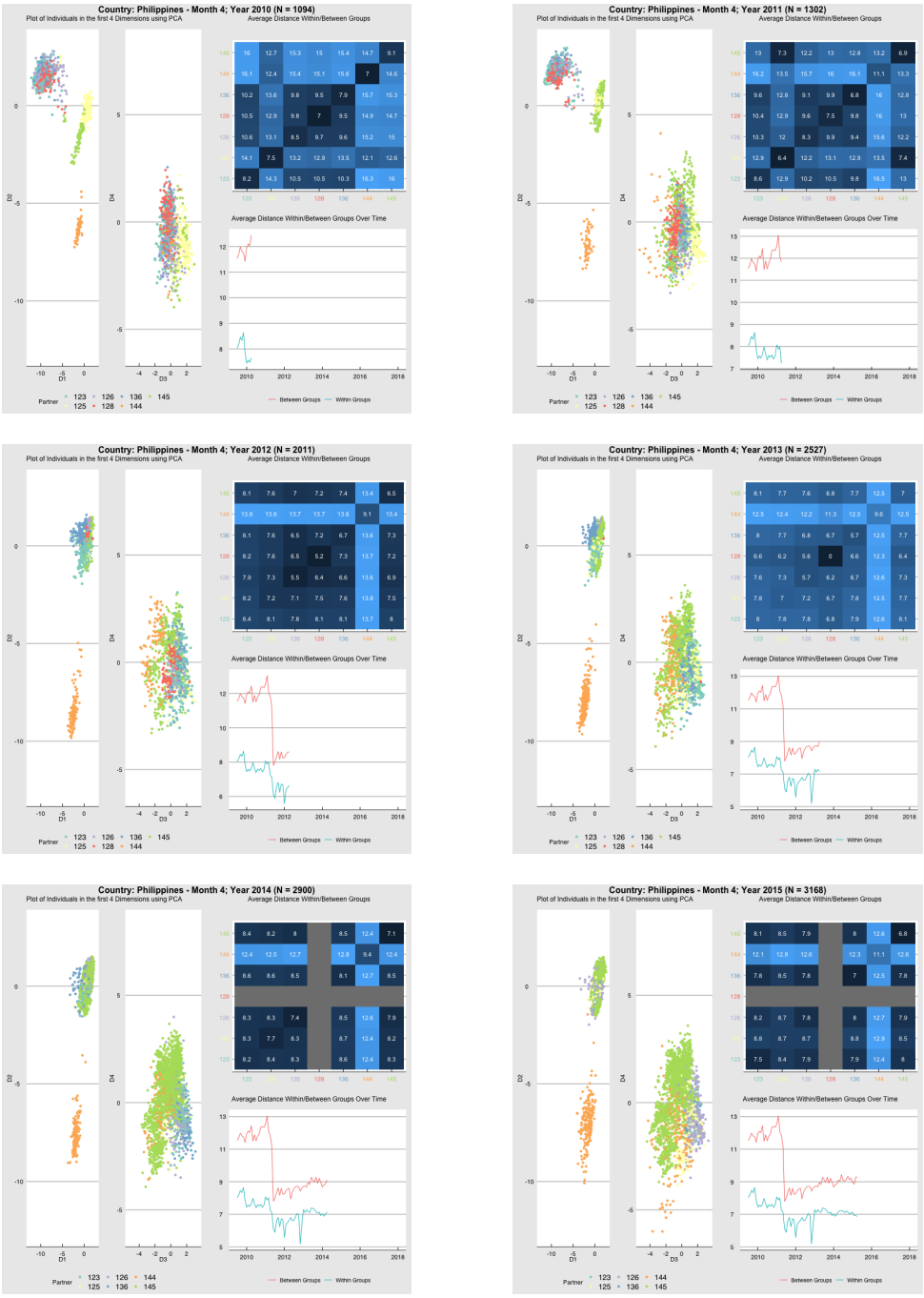


FIGURE 4.4: Evolution of Philippines Descriptions (2012 - 2017)

4.4 Authorship Attribution

In the previous section, we have seen how descriptions characteristics seem to be influenced by partner. In this section, we are willing to see how how good they can be distinguished.

To do so, a Machine Learning algorithm will be used. Support Vector Machines tend to perform good in this kind of problems; therefore we are

going to use two different methods: a linear Kernel and a radial one.

We have used a small portion of our dataset to perform this experiment. In this case, the subset of descriptions correspond to all those loans uploaded during the first quarter of 2011 (1st January 2011 - 31st March 2011), for Philipinnes, corresponding to 19505 observations.

We also want to know how (1) feeding the algorithm with more variables and (2) realizing dimensionality reduction alters the performance and the training time. To do so, we reduce the dimensionality by applying Principal Component Analysis and using the scores of the Principal Components instead of the Document-Term Matrix. In both cases, we'll be training the algorithm with different subsets of variables, always starting for the most informative ones and increasingly including the others. The sorting on the Document-Term Matrix is based on the total count of the variables: the variables with higher counts are considered those that contain more information.

We then report the Accuracy on the Test Set for every method. The original dataset is splitted on 70%Train and 30%Test. When Training, we implement 10-Fold Cross Validation.

4.4.1 Evaluation of Number of features and transformations to do: Results

The results are summarized on Test Data; in Figure **??** and Table 4.2. The outcome is clear: when using a Support Vector Machine, reducing the data shows a great improvement if a lower number of dimensions is used to feed the algorithm. Using the scores also reduces computational time. On the other hand, a linear kernel shows better accuracy then using the scores from PCA, whereas for the raw Term-Document Matrix, the radial kernel performs better. Regarding their computational time, a radial kernel is always more costly. All this is shown in Figure **??**. Another obvious and interesting outcome is the clear ability to distinguish and attribute ownership of every text. The confusion Matrix is presented in Table 4.2.

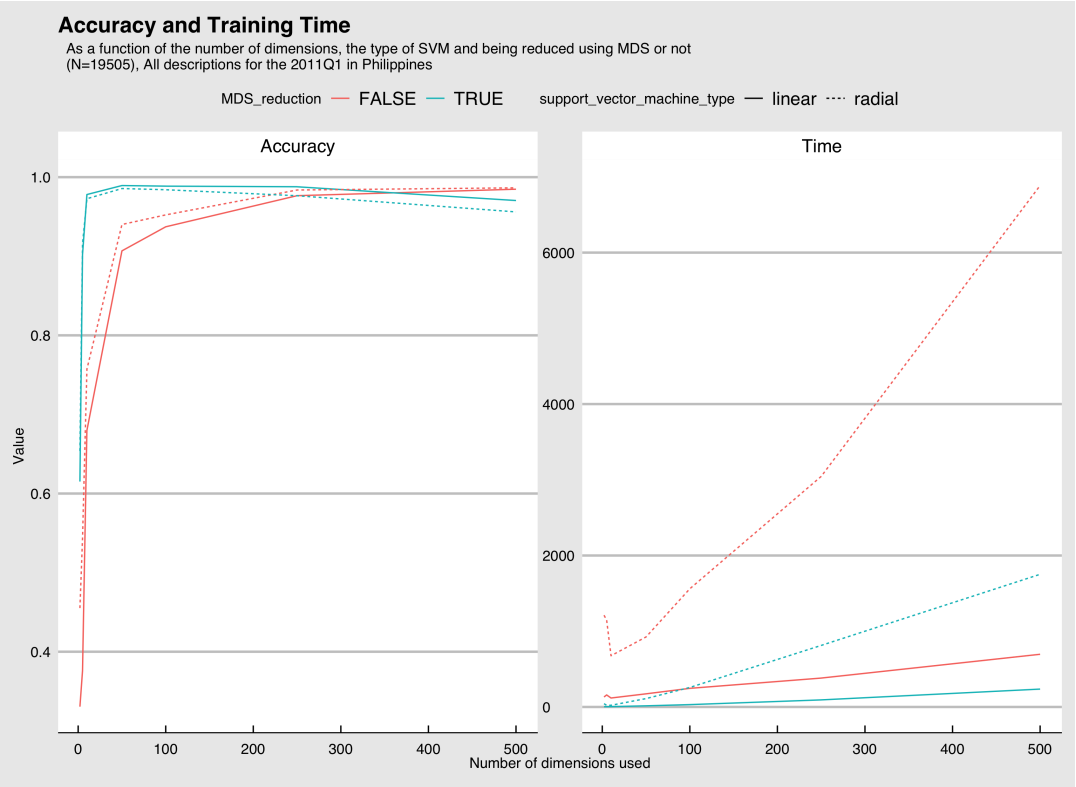


FIGURE 4.5: Evaluation of Number of features, transformations and Algorithms to classify Partners' Descriptions

TABLE 4.2: Confusion Matrix of the Test Data

	123	124	125	126	128	136	144	145
123	180	0	0	5	0	0	0	0
124	0	6	1	0	0	0	0	0
125	0	0	195	3	0	0	0	0
126	0	0	1	175	0	0	0	0
128	0	0	0	0	289	0	0	0
136	0	0	0	0	0	135	0	0
144	0	0	0	0	0	0	48	0
145	0	0	1	3	0	0	0	278

Chapter 5

Loans Images

5.1 Introduction

This chapter is focused on working on our borrower image hypothesis:

H6: Machine Learning can be used to extract expression labels on images.

H7: The borrowers face expression on the image has an impact on the loan performance.

In the past, it has already been attempted how graphical information (as in pictures) explained loan performance. An example is (Jenq, Pan, and Theseira, 2011), that aimed to investigate objective loan information, pictures and textual descriptions as determinants of individual charitable giving. In it, Research Assistants were asked to assess each photograph for qualities such as the borrower's appearance, age, gender, perceived honesty, and skin color on. Even though in this study multiple Research Assistants coded manually on the same picture to ensure consistency, the perception of a human being is not as stable as an Artificial Intelligence algorithm would be.

In the following chapter, we explore how, by using state of the art algorithms, facial expressions can be extracted from images. We finally seek if this facial expressions do have any impact on lender preferences.

5.2 Data Collection

In this thesis, we will make use of Google's Cloud Machine Learning Engine[1] and Microsoft's Azure Facial Recognition Software [2]. They have been chosen

[1] https://cloud.google.com/vision/
[2] https://azure.microsoft.com/en-us/services/cognitive-services/face/

because of their cutting edge technology and API access.

5.2.1 Limitations

Unfortunately, this is the section of the thesis that was less scalable. The bottleneck resided in obtaining the images as Kiva would block the client when realizing brute force requests. This made the data collection way slower. On the other hand, both APIs are not free. Both require credit card authorisation and even though free trials were used during this analysis, aiming for a very high amount would suppose monetary cost.

5.2.2 Working with a Subset

Because of the already explained limitations, we considered desirable to work with a subset of data. In order to reduce variance of the data due to non-interested variables, we aim to select a homogeneous subset. To do so, some restrictions are imposed in the original dataset, resulting in a sample of size 2696. The restrictions can be summarized on the Listing 5.1:

```
dt <- loans %>%
        # Country is Philippines
  filter(country_name=="Philippines",
        # Partner Id is 145
        partner_id==145) %>%
       # Posted Date is after 2016-03-01
  filter(to_date(posted_time)>='2016-03-01',
        # Posted Date is before 2016-04-01
        to_date(posted_time)<'2016-04-01') %>%
        # There is only ONE borrower; and is FEMALE
  filter(borrower_genders=="female",
        # There is only ONE borrower IN THE PICTURE
        borrower_pictured=="true",
        # The repayment interval is irregular
        repayment_interval=="irregular",
        # The distribution model is through field partner
        distribution_model=="field_partner",
        # The sector Name is either Agriculture, Food or Retail
        sector_name%in%c("Agriculture","Food","Retail"))
```

LISTING 5.1: R Code to select a subset

5.2.3 Scrapping the images

With Kiva not offering any method to obtain the image via API, web scrapping is used. The R library `rvest` becomes very handy to make the process less tedious. The process is as simple as accessing the different HTML nodes

of the webpage and then extracting the matching Regular Expression that contains an image. It is shown in Listing 5.2.

```
1
2  get_image_link <- function(x){
3
4    url <- paste0("https://www.kiva.org/lend/",x)
5    link <- read_html(url) %>%
6      html_nodes(xpath='//*[@id="main"]') %>%
7        html_nodes('div') %>%
8        html_nodes('div') %>%
9        html_nodes('div') %>%
10       html_nodes('figure') %>%
11       as.character() %>%
12       str_extract(pattern = "https://.*.jpg")
13   if(!identical(link,character(0))){
14     return(link)
15   } else {
16     return(NA)
17   }
18 }
```

LISTING 5.2: R Code to obtain a loan image

5.2.4 Google Cloud Machine Learning Engine

Among many of Google Cloud Platform Products, Google has a dedicated section to Artificial Intelligence. Inside Artificial Intelligece & Machine Learning Products, users can either train (by providing training data) or use a pre-trained model. The use case of this problem was to use the pre-trained model; because we did not have a labelled dataset.

The product used has been the **Cloud Vision API**, which "Integrates Google Vision features, including image labeling, face, logo, and landmark detection, optical character recognition (OCR), and detection of explicit content, into applications.".

The way to handle the requests has been through the build in library in R `RoogleVision` [3] by the use of POST methods. We then created a custom function where, by providing a link to a image, we would request the feature `FACE_DETECTION`.

Many results are returned, such as face coordinates of the `RIGHT_OF_LEFT_EYEBROW`, but the ones that were of interest were the face emotion labels.

The retrieved results of interest is a vector that contains, for every identified face, different variables: `joy, sorrow ,anger ,surprise ,underExposed ,blurred ,headwear` and a categorical ordinal value for every variable: `very`

[3] https://www.r-bloggers.com/google-vision-api-in-r-rooglevision/

unlikely, unlikely, possible, likely, very likely. The R Code is shown in the Listing 5.3.

```
1
2  # Reads a JSON file with the credentials.
3  # This credentials are not found on the public Github repository;
4  # as they were linked to my personal account with my credit card.
5  # The credentials can be downloaded on Google's Cloud Shell.
6  creds = fromJSON('credentials/credentials.json')
7
8  # sets the options from the credentials file
9  options("googleAuthR.client_id" = creds$installed$client_id)
10 options("googleAuthR.client_secret" = creds$installed$client_secret)
11
12 options("googleAuthR.scopes.selected" = c("https://www.googleapis.com/auth/
       cloud-platform"))
13
14 # Authorizes googleAuthR to access your Google user data
15 googleAuthR::gar_auth()
16
17 # Creates a custom function to retrieve the data
18 get_google_response <- function(x){
19   # calls the function
20   dt <- getGoogleVisionResponse(x,
21                                 feature = "FACE_DETECTION") [1,]
22   # establish two conditions to return NA in case of unvalid results
23   if(dt=="No features detected!"){
24     return(rep(NA,9))
25   } else if (ncol(dt)<9){
26     return(rep(NA,9))
27   # returns a vector with the desired output
28   } else {
29     res <- c(dt$detectionConfidence,
30              dt$landmarkingConfidence,
31              dt$joyLikelihood,
32              dt$sorrowLikelihood,
33              dt$angerLikelihood,
34              dt$surpriseLikelihood,
35              dt$underExposedLikelihood,
36              dt$blurredLikelihood,
37              dt$headwearLikelihood)
38     return(res)
39   }
40 }
```

LISTING 5.3: R Code to call the Google API and store the results

5.2.4.1 Output

As mentioned before, the original dataset for this section is a subset consisting of 2696 observations. Out of the 2696 requests to the Google Vision API, there has been 2448 (90.80%) valid responses.

However, based on the Output shown in Figure 5.1, there are some categories that are *almost* always labelled as VERY_UNLIKELY. Those are anger,

`blurred`, `surprise` ,`underExposed`. Due to their unexisting variance, they are dropped at this point from the analysis. Also, since the focus is on faces expressions, `headwear` is dropped at this point too.

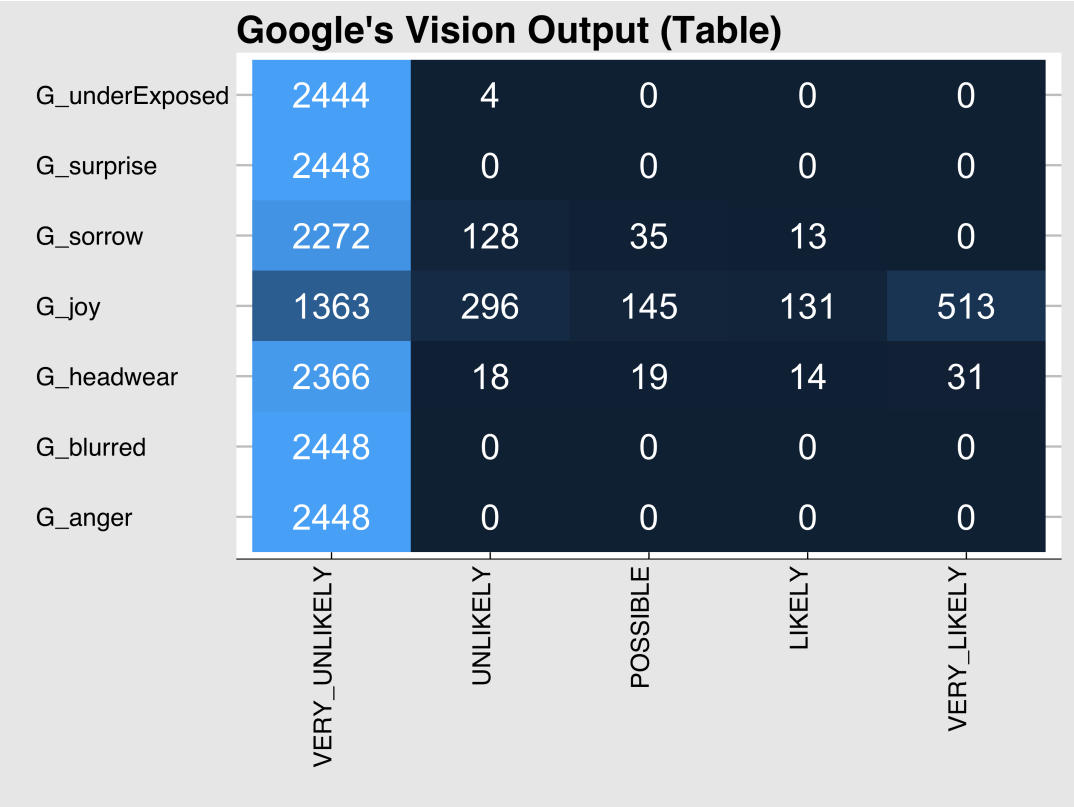


FIGURE 5.1: Google's Vision Output (as a Table)

5.2.4.2 From ordinal data to continuous data

Regarding Google Output,

> Results are categorized based on how likely they are to represent a match. The likelihood is determined by the number of matching elements a result contains. The Cloud Data Loss Prevention (DLP) API uses a bucketized representation of likelihood, which is intended to indicate how likely it is that a piece of data matches a given `InfoType`.

However, bucketizing the scores produces a loss of information. We are facing then, a Likert scale that for further simplicity in the analysis, we are willing to convert it to a continuous variable while minimizing information loss. We assume this *bucketized representation of likelihood* is done with buckets of the same distance. Then, the conversion to continuous data can be done through the following function (see Listing 5.4):

```
google_factor_to_numeric <- function(x){
  if (is.na(x)){
    return(NA)
  } else if (x=="VERY_UNLIKELY") {
    return(0)
  } else if (x=="UNLIKELY") {
    return(0.25)
  } else if (x=="POSSIBLE") {
    return(0.5)
  } else if (x=="LIKELY") {
    return(0.75)
  } else if (x=="VERY_LIKELY") {
    return(1)
  } else {
    return(NA)
  }
}
```

LISTING 5.4: R Code to convert the bucket values to numerical

5.2.5 Microsoft Azure

Regarding Microsoft, we have made use of its Facial Recognition Software: Face API. It offers a very similar product than Google, describing it as "Detect one or more human faces in an image and get back face rectangles for where in the image the faces are, along with face attributes which contain machine learning-based predictions of facial features. The face attribute features available are: Age, Emotion, Gender, Pose, Smile, and Facial Hair along with 27 landmarks for each face in the image."

In this case, there was no already existing library to make it friendly. However, by making use of the `httr` library in R, we managed to handle the requests by the use of POST methods. [4] by the use of POST methods. See Listing 5.5 for R Code reference. Exactly the same as for Google, we explicited our query to `returnFaceAttributes`, only aiming at `emotion`. Other possibilities within `returnFaceAttributes` included `age,gender,hair,makeup,accessories`. Regarding the output, Azure returns a numeric value for the following categories for every identified face: `anger, contempt, disgust, fear, happiness, neutral, sadness, surprise`. The sum of this numeric values for every picture object equals to one.

```

end.point <- "westeurope.api.cognitive.microsoft.com"
key1 <- "" #personal and not shareable

get_azure_response <- function(x){
```

[4] https://www.r-bloggers.com/using-microsofts-azure-face-api-to-analyze-videos-in-r/

```
6
7   # does the POST method
8   # inside the URL the requests are specified:
9   # returnFaceId=false
10  # returnFaceLandmarks=false
11  # returnFaceAttributes=emotion
12
13  res <- httr::POST(url = "https://westeurope.api.cognitive.microsoft.com/face/
        v1.0/detect?returnFaceId=false&returnFaceLandmarks=false&
        returnFaceAttributes=emotion",
14              body = paste0('{"url":"',x,'"}'),
15              add_headers(.headers = c("Ocp-Apim-Subscription-Key" = key1)))
16
17  # res is an object of category "response".
18  # we aim to access its content.
19
20  # check for valid results
21  if(length(httr::content(res))>0){
22  # return valid results
23    return(unlist(httr::content(res)[[1]]$faceAttributes$emotion,
24                  use.names=FALSE))
25  } else {
26  # returns NA
27    return(rep(NA,8))
28  }
29 }
```

LISTING 5.5: R Code to call the Google API and store the results

5.2.5.1 Output

Regarding Microsoft's Azure output, valid responses were in 2306 (85.53%) of the cases, 5.24p.p. lower than Google's Vision. As it can be seen in Figure 5.2, most of the scores are assigned to `happiness`, `neutral`, remaining low scores for the other variables. However, its responses register more variability in the whole set of variables. None of them is dropped and all the variables are used in the following sections.

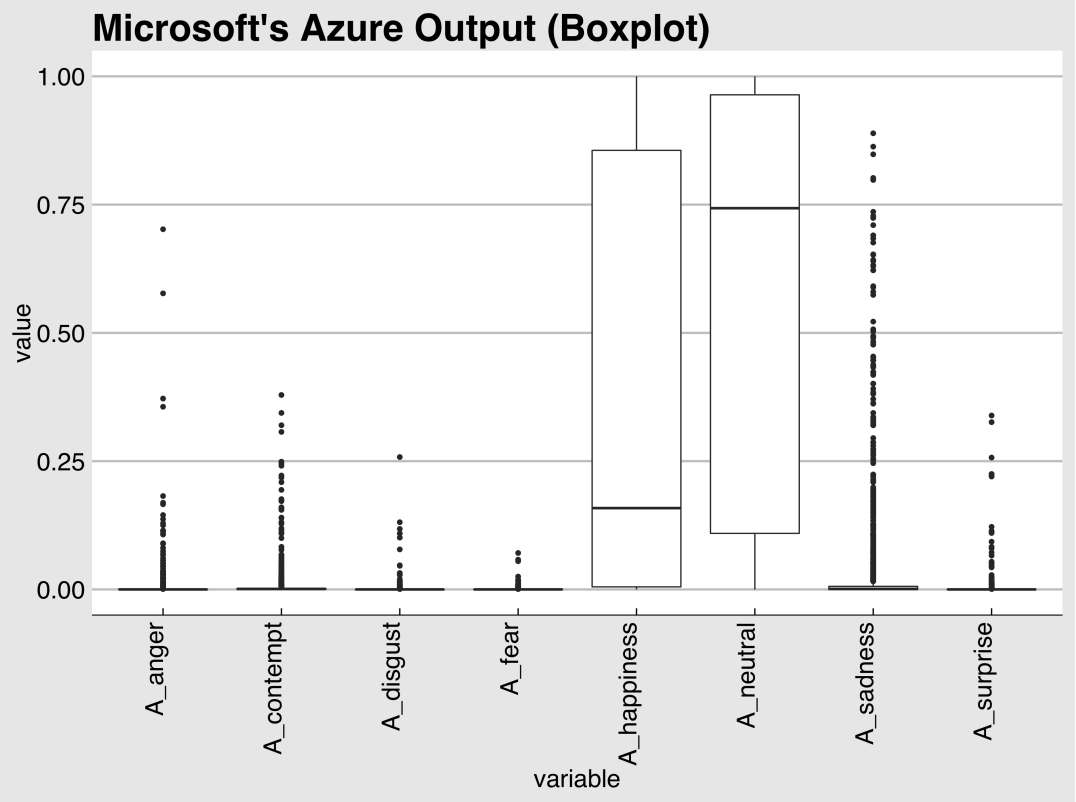


FIGURE 5.2: Microsoft's Azure Output (as Boxplot)

5.2.6 Example Visual Output

In order to give a visual example of the results, please see Figures 5.3 and 5.4.

FIGURE 5.3: Loan 1033283: Image and Output

FIGURE 5.4: Loan 1038440: Image and Output

5.3 Data Modelling

5.3.1 Exploratory Factor Analysis

Now that we have a set measured variables coming from two different sources, our goals are:

- **determine underlying factors/constructs**, so that we find a meaningful way to group the variables
- **understand the relationship between the variables**, as we are expecting high correlations between the two data sources.

For the type of data and our goals, the most appropiate technique to use is **Exploratory Factor Analysis (EDA)**.

In Figure 5.5 the variables Loadings are plotted in the first two dimensions. On the other hand, the Loadings are summarized in Table 5.1. The conclusions that can be extracted are:

- The first dimension explains XX of the variance. It is mostly represented by three variables: `A_happiness`, `G_joy` (that are highly positive correlated) and `A_neutral`, highly negative correlated to the previous two. This would support the sixth hypothesis:

  > **H6:** Machine Learning can be used to extract expression labels on images.

  Due to the fact that both APIs are able to distinguish between high levels of joy and low ones.

- The second dimension explains XX of the variance. It is highly influenced by `A_Anger`, `A_Disgust` and `G_sorrow`.

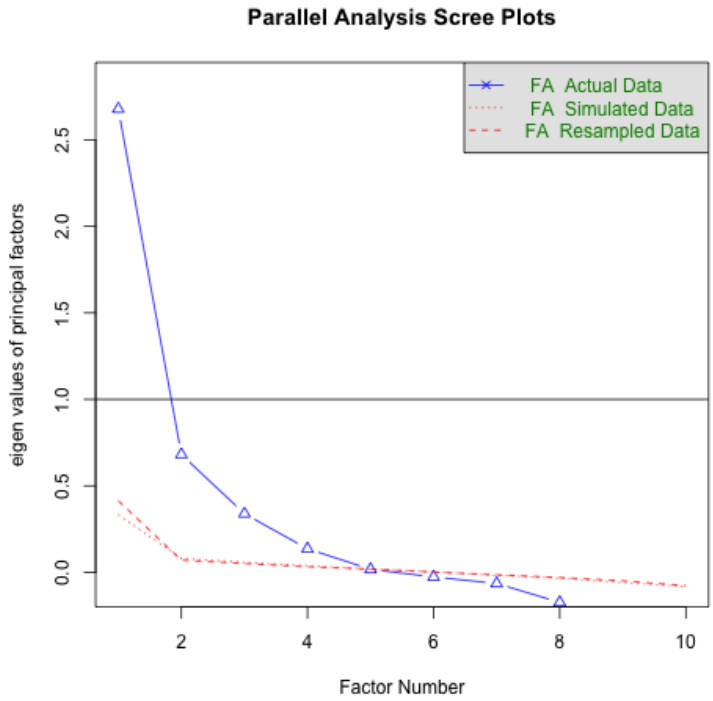


FIGURE 5.5: Factor Analysis: Variable Plot

	Factor1	Factor2	Factor3	Factor4
A_anger	-0.05	1.00	0.01	0.05
A_contempt	-0.03	0.04	0.06	0.99
A_disgust	0.01	0.57	0.08	0.02
A_fear	0.00	0.13	0.15	-0.01
A_happiness	0.97	-0.03	-0.21	-0.07
A_neutral	-1.00	-0.05	-0.03	0.01
A_sadness	0.02	-0.01	1.00	-0.03
A_surprise	-0.06	0.04	0.00	-0.02
G_joy	0.82	0.01	-0.10	-0.05
G_sorrow	-0.15	0.02	0.30	0.07

TABLE 5.1: Loadings of the Exploratory Factor Analysis

5.3.2 Multiple Linear Regression

After realizing the **EDA** and gathering some insights on the relationship of the variables obtained from the Image Recognition APIs, we are willing to work on our last hypothesis:

H7: The borrowers face expression on the image has an impact on the loan performance.

To do so, the methodology used will be a Multple Linear Regression **MLR**. In the **MLR**, we aim to explain a dependent variable as a function of several explanatory variables. Our dependent variable will be `time_to_fund`. The `time_to_fund` for a given loan is the difference between the variables `posted_time` and `raised_time`. The `time_to_fund` median is 90 hours. We are going to propose several models. The variables that are going to be common in all of them are:

- `is_monday`: After realizing Exploratory Data Analysis, we observed that those loans whose `posted_time` is a Monday in the time zone *America - Los_Angeles* (as Kiva HQ are in San Francisco, California) have significantly shorter funding time.
- `is_retail`: Out of the 3 different `sector_name` used (Agricultur, Food and Retail), Retail had a significantly longer funding time.
- `loan_amount`: We expect that a higher loan amount causes the funding time to be longer.

Willing to understand the impact of Face Labels, we'll suggest two different sets of variables.

- The first set, named **Manual_Scores** will be a manual selection of some scores obtained. In this case, interpretability of the variables will be easier. The variables created will be:

  `happiness`. It is a construction of `G_joy+A_happiness-A_neutral`.

  `negative_others`. It is a construction of `A_anger + A_disgust + G_sorrow + A_sadness + A_contempt + A_fear`.

- The second set, named **Factor_Scores** will contain the individual scores resulting by the Exploratory Factor Analysis. The first three dimensions are included:

  `FA_1` First Dimension Individual Scores of the **EDA**

  `FA_2` Second Dimension Individual Scores of the **EDA**

  `FA_3` Third Dimension Individual Scores of the **EDA**

On the other hand, we will also present the two alternatives as the response variable: `time_to_fund` and `log(time_to_fund)`. Three different models are specified:

- Multiple Linear Regression Model

- Multiple Linear Regression Model with Dependent Variable transformed (log-linear Model)
- Logistic Model

5.3.3 Results

After running all the models, the AIC and BIC show as great candidates with the log-linear Model. The **Manual_Scores** shows a signficant coefficient for `happiness` but `negative_others` is not significant. The **Factor_Scores** shows a signficant coefficient for `FA_1`, `FA_2` but `FA_3` is not significant. We proceed to evaluate that the assumptions of the linear model are held:

- The mean of the residuals is 0.
- Homoscedasticity of residuals or equal variance
- No autocorrelation of residuals
- Normality of residuals

Apart from normality of residuals, all of them are held. The normality of residuals, as seen in the QQ-plot, is weakly violated, being then not a major concern.

Regarding the interpretation of the coefficients, being a Log-Level Regression, $ln(y) = \beta_0 + \beta_1 * ln(x_1) + \epsilon$, a increase of x by one unit, expects an increase of $\beta_1 * 100\%$ units of y.

As in model (2), the coefficient for `happiness` is -0.046. Being the difference of the 75th and 25th percentiles of `happiness` $P_{75} - P_{25} = 2.19$, the difference of a given image to be in the 75th percentile and the 25th percentile modifies the loan funding time by more than 10%. Model (5) estimates the coefficient of the first dimension of EFA to be -0.034 (significant with $\alpha = 0.05$) and the coefficient of the second dimension of EFA to be -0.034 (not significant with $\alpha = 0.05$, but significant at $\alpha = 0.1$).

5.4 Further Work

The findings motivate the increase of sample size in order to get more robust results. Finding that either both the first and second dimensions of EFA impact the performance of a loan is surprising.

On the other hand, other variables could be used, such as the resolution of the picture. The interaction of the face labels with the sentiment analysis of the description, working on the hypothesis of "similar images and description reduce the funding time of the loan". The hypothesis would be that when the image and the description go in the same direction (both happy or both sad), the loan performs better but when they go into opposite (one happy and the other sad), it harms loan performance.

TABLE 5.2: Image Performance: Regression Summary

	Dependent variable:					
	(time_to_fund)	log(time_to_fund)	(time_to_fund)		log(time_to_fund)	(time_to_fund)
	OLS	*OLS*	*normal*	*OLS*	*OLS*	*normal*
	(1)	(2)	(3)	(4)	(5)	(6)
loan_amount	0.355***	0.002***	0.355***	0.355***	0.002***	0.355***
	(0.032)	(0.0002)	(0.032)	(0.032)	(0.0002)	(0.032)
is_monday	−29.795***	−0.316***	−29.795***	−29.514***	−0.314***	−29.514***
	(8.444)	(0.046)	(8.444)	(8.449)	(0.046)	(8.449)
is_retail	39.378***	0.180***	39.378***	39.358***	0.180***	39.358***
	(6.708)	(0.036)	(6.708)	(6.714)	(0.036)	(6.714)
happiness	−9.836***	−0.046***	−9.836***			
	(2.908)	(0.016)	(2.908)			
negative_others	−1.935	−0.095	−1.935			
	(20.190)	(0.109)	(20.190)			
FA_D1				−7.961***	−0.034**	−7.961***
				(2.504)	(0.014)	(2.504)
FA_D2				−4.470	−0.034*	−4.470
				(3.472)	(0.019)	(3.472)
FA_D3				0.391	−0.005	0.391
				(3.957)	(0.021)	(3.957)
Constant	24.226**	4.096***	24.226**	22.912**	4.086***	22.912**
	(10.442)	(0.056)	(10.442)	(10.421)	(0.056)	(10.421)
AIC	24055.7	4404.2	24055.7	24057.7	4405.5	24057.7
BIC	24094.5	4443	24094.5	24102.1	4449.8	24102.1
Observations	1,882	1,882	1,882	1,882	1,882	1,882
R^2	0.086	0.077		0.086	0.078	
Adjusted R^2	0.084	0.075		0.083	0.075	
Log Likelihood	−12,021.860		−12,021.860	−12,021.870		−12,021.870
Akaike Inf. Crit.			24,055.720			24,057.740
Residual Std. Error	144.018 (df = 1876)	0.778 (df = 1876)		144.057 (df = 1875)	0.778 (df = 1875)	
F Statistic	35.281*** (df = 5; 1876)	31.451*** (df = 5; 1876)		29.382*** (df = 6; 1875)	26.328*** (df = 6; 1875)	
Note:						*p0.1; **p0.05; ***p0.01

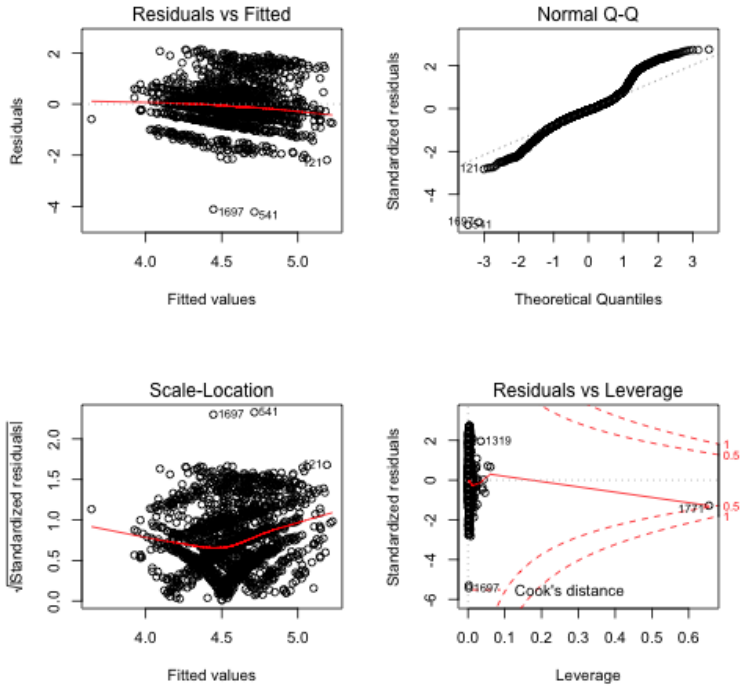


FIGURE 5.6: Validation of log-linear model

Chapter 6

Conclusions

In this Thesis, we have been focusing on solving completely different problems but with a common topic.

In Chapter 2, Literature Review is performed and Opportunities are found. Clear Hypothesis and Goals are stated for this Thesis.

In Chapter 3, we focus on the Lender-Borrower relationship. We explore how Female Borrowers receive more funding from Females, and how this differers by Sector. However, due to low interpretability we do not extract valid conclusions. We then provide an approach to describe the categorical relationships on Occupation and Country, rejecting the Null Hypothesis of the χ^2 test of Independence: both Countries and Occupations show strong deviations and therefore associations. When using Correspondence Analysis, it has not been possible to identify *similar* lending behaviour for *similar* countries neither similar lending behaviour for *similar* professions.

In Chapter 4, we focus on partners descriptions. We explore how different methods (Metric and Non-Metric Multidimensional Scaling) with different distances (Jaccard, Cosine and Euclidean) are explored to provide a visual representation of text. We finally provide a Framework to evaluate the evolution of descriptions over time.

Finally, we showcase how, by using a Suport Vector Machine, partner authorship can be *almost* perfectly attributed.

In Chapter 5, we have used Machine Learning to extract Face Expressions from Images. We have later seen how the happiness of a face is properly identified; finding difficulty in identifying other expressions. This is also shown when seeking the importance of Face Expressions in the Loan success: results show that loans whose Face Images have high scores of Happiness fund faster (easily 10% faster!), but it is not relevant with other emotions. As further actions, considering the interactions with a sentiment output from the descriptions is suggested.

Last but not least, another goal has been archieved. This Thesis is public and completely available for reproducibility, containing all the code and files at https://github.com/mvalenti12/TFM.

Bibliography

Alegre, Ines and Melina Moleskis (2017). "Crowdfunding: A Review and Research Agenda". In: *Ssrn* 3. DOI: 10.2139/ssrn.2900921.

Allison, Thomas H. et al. (2015). "Crowdfunding in a prosocial microlending environment: Examining the role of intrinsic versus extrinsic cues". In: *Entrepreneurship: Theory and Practice* 39.1, pp. 53–73. ISSN: 15406520. DOI: 10.1111/etap.12108.

Burtch, Gordon, Anindya Ghose, and Sunil Wattal (2013). "Cultural Differences and Geography as Determinants of Online Pro-Social Lending". In: *Ssrn* \. ISSN: 1556-5068. DOI: 10.2139/ssrn.2271298.

Canela, Miguel Ángel (2017). "Working Paper WP-1172: Crowdfunding in Africa: An Empirical Study of Kiva Platform Users in Sub-Saharan Africa". In: 3.May.

Galak, Jeff, Deborah A. Small, and Andrew T. Stephen (2010). "Micro-Finance Decision Making: A Field Study of Prosocial Lending". In: *Ssrn*, pp. 1–55. ISSN: 0022-2437. DOI: 10.2139/ssrn.1634949. arXiv: 2794335.

Greenberg, J. and E. Mollick (2015). ""Leaning In or Leaning On? Gender, Homophily, and Activism in Crowdfunding"". In: *Academy of Management Proceedings* 2015.1, pp. 18365–18365. ISSN: 0065-0668. DOI: 10.5465/AMBPP.2015.18365abstract. URL: http://proceedings.aom.org/cgi/doi/10.5465/AMBPP.2015.18365abstract.

Jenq, Christina, Jessica Pan, and Walter Theseira (2011). "What Do Donors Discriminate On? Evidence From Kiva.org". In: May.

Kiva. *Kiva Website | About Kiva*. URL: https://www.kiva.org/about (visited on 09/17/2018).

Lin, Mingfeng and Siva Viswanathan (2013). "Home Bias in Online Investments: An Empirical Study of an Online Crowdfunding Market". In: *Ssrn* 2013. ISSN: 0025-1909. DOI: 10.2139/ssrn.2219546.

Moleskis, Melina and Ines Alegre (2018). "Crowdfunding: A Short Past and Long Future". In: *Ssrn*. DOI: 10.2139/ssrn.3163006.

Moleskis, Melina and Miguel Ángel Canela (2016). "Crowdfunding Success: The Case of Kiva.org". In: 3.

Nagpaul, P S (1999). "Correspondence analysis". In: *Guide to advanced data analysis using IDAMS software.* Pp. 1–20.

Pope, Devin G and Justin R Sydnor (2011). "What's in a Picture ? Evidence of Discrimination from Prosper.com". In: *The Journal of Human Resources* 46.1, pp. 53–92. ISSN: 1548-8004. DOI: 10.1353/jhr.2011.0025.

Zeileis, Achim, David Meyer, and Kurt Hornik (2007). "Residual-based shadings for visualizing (conditional) independence". In: *Journal of Computational and Graphical Statistics* 16.3, pp. 507–525. ISSN: 10618600. DOI: 10.1198/106186007X237856.