ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Department of Statistical Sciences "Paolo Fortunati"

Second Cycle Degree in Statistical Sciences

Social Demography

Language and Beyond: How Linguistic and Socio-Economic Integration Affects Employment Among Immigrants in Italy

Sona Yavrumyan and Lavinia Sposetti

Academic Year 2023/2024

Contents

Abstract

This research paper analyzes the integration of immigrants in Italy, focusing on linguistic proficiency and socio-economic key indicators. Using the SCIF dataset from Istat[1], we statistically analyze the impact of a multitude of variables on employment outcomes, emphasizing gender, age, and ethnic background. Our findings reveal significant differences in labour market integration across different ethnic backgrounds, and gender and age disparities, with younger female immigrants facing substantial employment challenges. We discovered that, household size significantly affects employment, as larger families face greater difficulties securing employment. Immigrants with East and South-East Asian ethnicity, show higher employment rates compared to others. European-origin female immigrants exhibit higher employment probabilities, while Arabic-speaking women showcase the lowest rates across all groups. We recommend tailored language and job training programs, gender-specific support, and expanded educational opportunities to enhance immigrant integration.

1 Introduction

The integration of immigrants into host societies has become a central focus of sociological, demographic, and economic research, particularly in countries like Italy that experience significant migration flows due to their strategic geographical position. Since the 1990s, Italy's national statistical agency, Istat, has systematically included variables related to citizenship in their data collections, aiming to capture detailed insights into the foreign population's presence and migration trends.

Contrary to the common perception that Southern European countries like Italy lag in migrant integration, recent analyses suggest a more nuanced reality. Using the MIPEX (Migrant Integration Policy Index) dataset[2], which covers migrant integration policies across OECD countries, it becomes evident that Southern European countries, including Italy, do not conform to a single model of integration and are not necessarily less developed in their policies compared to their Northern European counterparts. Italy has made strides in progress in various areas of immigrant integration over the past few decades.[3]

Immigrants primarily integrated into Italian society through labour markets, taking on manual and low-skilled jobs that Italians abandoned, especially in informal sectors like small enterprises, construction, tourism, agriculture, and domestic services. This need for labour has resulted in a type of "subordinated integration," where immigrants perform heavy, precarious, and low-paid jobs[4].

This paper seeks to explore the multifaceted nature of immigrant integration in Italy, with a particular emphasis on linguistic aspects (both in terms of the proficiency of immigrants in the Italian language and the use of their mother tongues) and key socio-economic dimensions. By utilizing the SCIF dataset, we will conduct an exploratory analysis to derive insights about the immigrant population and perform statistical analysis to identify which variables significantly affect employment outcomes among immigrants and why that could be the case in the general outlook of immigration integration policies and outcomes.

This study will explore the unique challenges faced by female immigrants, including potential gender biases and the additional responsibilities often borne by women. A specific focus will be placed on how age and ethnicity interact with influences on their integration into the labour market. Our findings from this study will contribute to a deeper understanding of the challenges and opportunities associated with immigrant integration in Italy, especially for integration into the labour market.

2 Framework in Current Migration Integration Studies

To understand general integration trends, we utilized the existing Migrant Integration Policy Index (MIPEX), which provided a comprehensive overview of Italy's performance relative to other countries. It is important to note in the framework of integration comparisons, that Italy does not follow a single national model of integration but instead has multiple local models. This decentralized approach focuses on the importance of local identity and the varied experiences of integration across different regions, so this could potentially introduce some slight difficulties in comparing Italy to other countries as a whole. Regardless, MIPEX is a robust and widely acknowledged instrument designed to evaluate and compare integration policies across various nations, particularly within the OECD. It systematically assesses policies in eight critical domains of integration: labour market mobility, family reunion, education, political participation, permanent residence, access to nationality, anti-discrimination, and health.

In the context of labour market mobility, MIPEX classifies Italy's policies as moderately favorable. Non-EU citizens in Italy have access to employment and self-employment opportunities, which supports their economic integration. However, these policies frequently lack targeted support mechanisms necessary for non-EU immigrants to secure employment that corresponds with their qualifications and skills. While Italy has made significant strides in providing labour market access, additional measures are imperative to enhance job matching and career advancement for immigrants.

To get a more comprehensive idea, we aimed to compare (Figure 1) Italy's immigrant integration policies to those of the broader EU (EU15 - the 15 pre-2004 EU member states), given their similarities, and also to a leading country in the implementation of immigraiton integration

policies, Canada. Thus, we chose these three regions for our analysis. The bar chart provides a comparative overview focusing on Overall Score, Labour Market Mobility, Education, and Anti-Discrimination. While Canada emerges as the leader with the highest overall scores, indicating a robust framework for immigrant integration, Italy's scores are not significantly lower than the EU15 average. This suggests that, contrary to perceptions of underperformance in integration initiatives, Italy's efforts are relatively aligned with the broader European standard. One area for potential improvement is enhancing access to higher education for new immigrant students, as this can significantly impact their employability, career choices, and overall life trajectory.

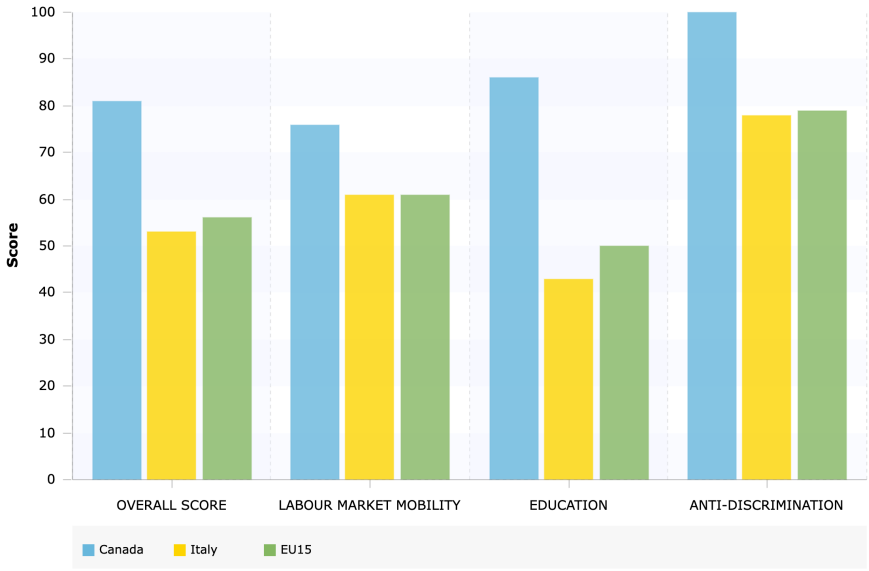

Figure 1: MIPEX scores in Italy, Canada and EU15 in 2012 across three categories

3 Data Under Study

Istat, in collaboration with the Ministry of Interior of Italy has conducted various research projects, aimed to deepen the understanding of migrants' integration, guided by principles laid out in documents such as the 2010 Zaragoza Declaration[5]. Istat has been working towards maximizing the use of administrative data to develop a system of integrated registers and has conducted two significant surveys focusing on integration issues, one of which (SCIF - SOCIAL CONDITION AND INTEGRATION OF FOREIGN CITIZENS) will be used for integration analysis in this paper.

The survey's methodology[6] included a two-stage selection process involving 833 municipalities and 9,553 households, ultimately interviewing a total of 25,326 individuals. The sampling technique employed was a stratified random approach, with municipalities selected based on the size of their foreign resident populations, particularly focusing on the 15 most represented nationalities. The breadth of the SCIF survey is extensive, capturing data across various aspects such as

family composition, educational attainment, migration history, employment status, experiences of discrimination, health conditions, societal integration, personal security, and housing conditions. The SCIF survey employed the Computer Assisted Personal Interviewing (CAPI) method, which effectively managed different questionnaire paths for various population segments. The presence of interviewers was crucial in establishing initial contact, enhancing respondent cooperation, and ensuring data consistency and completeness. To aid communication and comprehension, the questionnaire was translated into major foreign languages. For individuals under 14 and temporarily absent respondents, proxy interviews were conducted, with responses provided by a parent or adult family member.

In our analysis of the SCIF survey data, we specifically utilized a subset of the dataset that was most relevant to our research objectives, focusing on employment status as the dependent variable. The recoding of this variable is detailed in Figures 2 and 3. Alongside employment status, we included key covariates such as age, sex, education, ethnic background, and various language-related variables, to comprehensively assess factors influencing immigrant employment in Italy.

To facilitate our statistical analysis and enhance interpretability, we transformed all numerical variables into categorical ones. For example, the age variable was categorized into different age groups. Additionally, we transformed categorical variables with extensive categories, such as mother tongues, into broader groups. Language related variables which seemed to display similar information, such that they showed multicolinearity were aggregated together into one variable.

Any records with missing values for our dependent variable were excluded to ensure the accuracy of our analysis. For covariates with missing values, we applied mean imputation for numerical variables and mode imputation for categorical variables. This imputation strategy helped maintain a robust dataset.

		Freq.	Percent	Cum.
Employed -	Employed	9,185	54.46	54.46
	collaboration work	257	1.52	55.99
	Self-employment	1,245	7.38	63.37
	Seeking employment	1,517	8.99	72.36
	Inactive	4,661	27.64	100.00
	Total	16,865	100.00	

Figure 2: Original dependent variable - Employment status in 5 categories

	Freq.	Percent	Cum.
Not working	6,178	36.63	36.63
Working	10,687	63.37	100.00
Total	16,865	100.00	

Figure 3: Recoded dependent variable - Employment status in 2 categories

4 Exploratory Descriptive Analysis

To get an idea of the immigrant population, we initially explored general trends using the 2011 Census from Istat[7]. This dataset provides final data on the resident population by sex,

age, and citizenship, which are crucial for understanding the composition and dynamics of the Italian population. By combining these sources of information, we constructed a bar chart (Figure 4) comparing the percentage of the Italian population with that of immigrants across various age groups. Our analysis reveals that the immigrant population is predominantly young, particularly in the 0-35 age group, highlighting significant demographic differences between immigrants and the Italian population. For example, immigrants constitute about 8.7% in the 25-30 age group compared to 5.5% for Italians. As age increases, the proportion of immigrants decreases significantly, with only 0.1% of those aged 85-90 being immigrants, compared to 3.3% of Italians, suggesting older immigrants might return to their countries of origin or that significant immigration is a recent phenomenon. The immigrant population peaks at 11.3% in the 35-40 and 40-45 age brackets, indicating many are in their prime working ages, which may impact the labour market and economic contributions. This demographic distribution could imply a greater need for educational and other child-related services for immigrants, while the older Italian demographic suggests increased pressure on healthcare and retirement systems, underlining the importance of integrating younger immigrants through education and social programs to maximize their contributions.

In general, studies show that education plays a crucial role in integrating second-generation immigrants. However, the Italian education system has struggled to effectively support these children, resulting in significant school failures compared to native students. Despite various memorandums and initiatives aimed at promoting intercultural education, challenges remain in providing adequate support and ensuring successful integration through the education system .

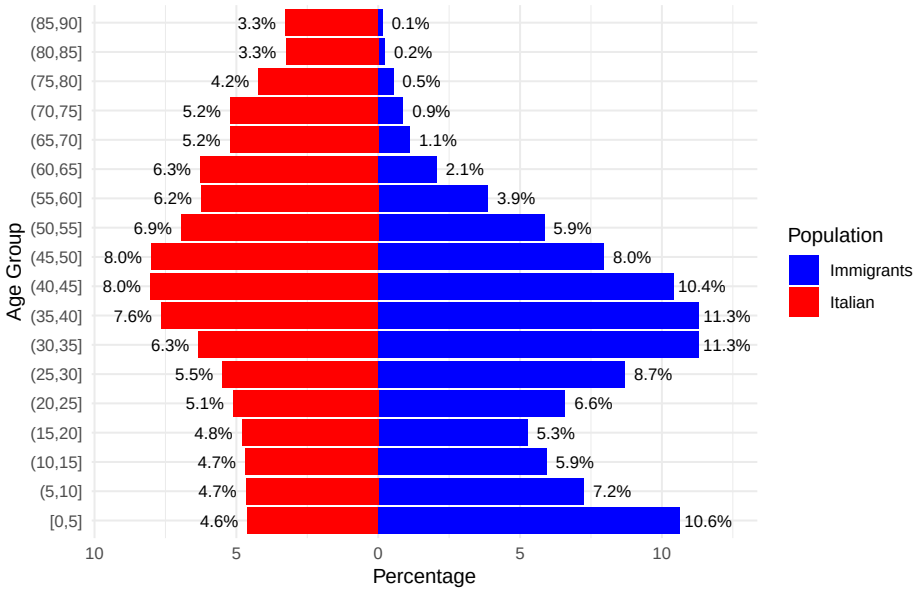

Figure 4: Population Pyramid for the Italian Population and immigrants in 2012

After the data manipulation and cleaning, we wanted to understand how well the immigrant population has mastered the Italian language in terms of comprehension by the natives, as

language is of key interest to our research objective. The bar chart in Figure 5 illustrates proficiency levels in reading (red) and writing (blue) Italian across four categories. Notably, a significant proportion of respondents report no proficiency ("Not at all"), with 44.0% in reading and 35.6% in writing, indicating substantial language barriers. In contrast, smaller segments claim high proficiency ("A lot"), with 7.8% in reading and 10.4% in writing. This distribution suggests challenges in language acquisition among the population, potentially impacting their integration and access to opportunities within the Italian society. The disparity underscores the need for targeted educational programs to enhance both skills concurrently.

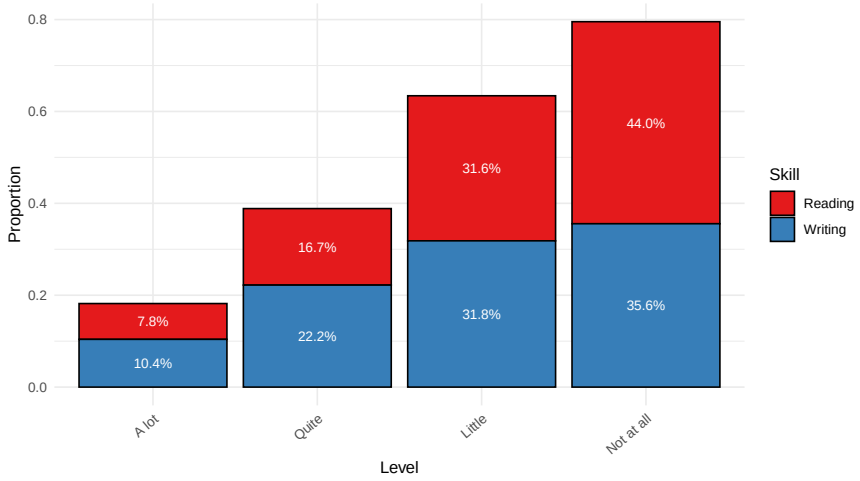

Figure 5: Reading and Writing Knowledge of the Italian Language in the Immigrant Population

The bar chart in Figure 6 visualizes the proficiency of Italian language understanding across different mother tongues. Categorizing immigrants' ethnicity based on their mother tongue rather than their country of origin provided an interesting outlook. This approach allowed us to capture cultural similarities more effectively, grouping together individuals with similar linguistic and cultural backgrounds (such as Serbians and Montenegrins) and avoiding the redundancy of having too many similar categories. In the plot, the level of spoken Italian is rated on four levels of understanding: "Quite well," "Moderately," "Poorly," and "Not at all." The chart shows significant linguistic diversity, with Chinese speakers reportedly being understood "Quite well" by Italians most frequently, whereas most Central-Eastern Europeans have a similar distribution across the categories, with a larger proportion indicating "Not at all." This pattern reflects somewhat varying levels of linguistic integration among immigrant communities in Italy, suggesting potential barriers in communication which could affect social integration, employment, and access to services. We can note that, overall, there is a need for tailored language education programs that cater specifically to the linguistic backgrounds of immigrant populations.

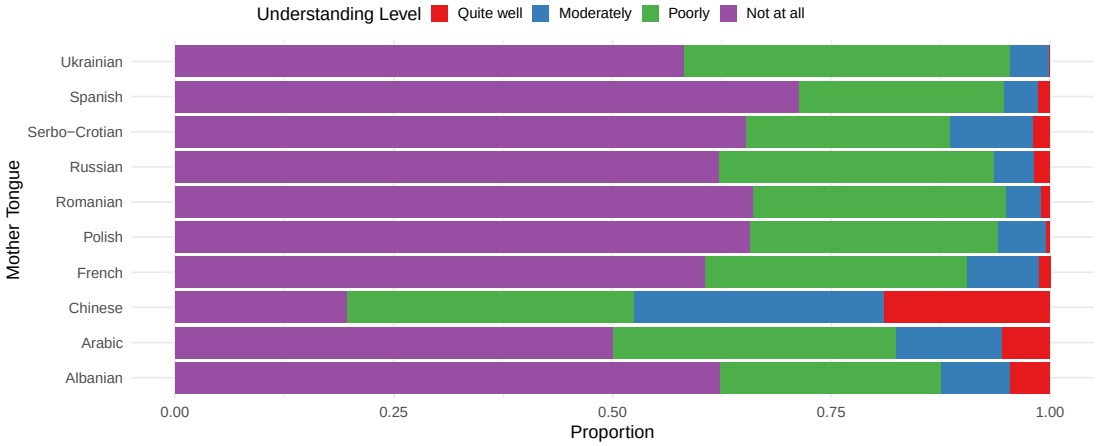

Figure 6: Level of Spoken Italian (How Well Can an Italian Understand You)

5 Multivariate analysis

In our study, we conducted an extensive multivariate analysis after carefully preparing the dataset, which included data cleaning, subsetting, recoding, and factoring variables, along with necessary imputation of missing values. We fitted two types of statistical models to improve the robustness of the results. The first model was a Generalized Linear Model (GLM) with a binomial distribution, suitable for modeling binary outcomes(which in our case is the status of employment variable with Yes/No categories). The second model was a LASSO (Least Absolute Shrinkage and Selection Operator) regression[8], which incorporates a regularization penalty to reduce overfitting and address multicollinearity by shrinking less influential predictors' coefficients towards zero.

We assessed the influence of different variables using a standardized measure of "Variable Importance"[9], which ranges from 0 to 100, with higher values indicating greater influence on the model outcome. This measure was crucial for identifying which variables most significantly affect employment status in our analysis. For the GLM, variable importance was determined based on the absolute values of the t-statistics for each model parameter, while for the LASSO model, it was directly based on the absolute values of the coefficients from the tuned model.

These analyses enabled us to visualize and compare the significance and impact (positive or negative) of various predictors across both models, providing insights into which factors are most critical in determining employment status.

Figure 7 shows the variable importance chart (for the top 30 most influential variables) for the GLM model. From the chart, it is evident that "Gender: Female" and "Workers: Aged 31-64" are the top predictors, suggesting that gender and age significantly impact employment status. This aligns well with existing literature that emphasizes the influence of demographic factors on labour market participation.

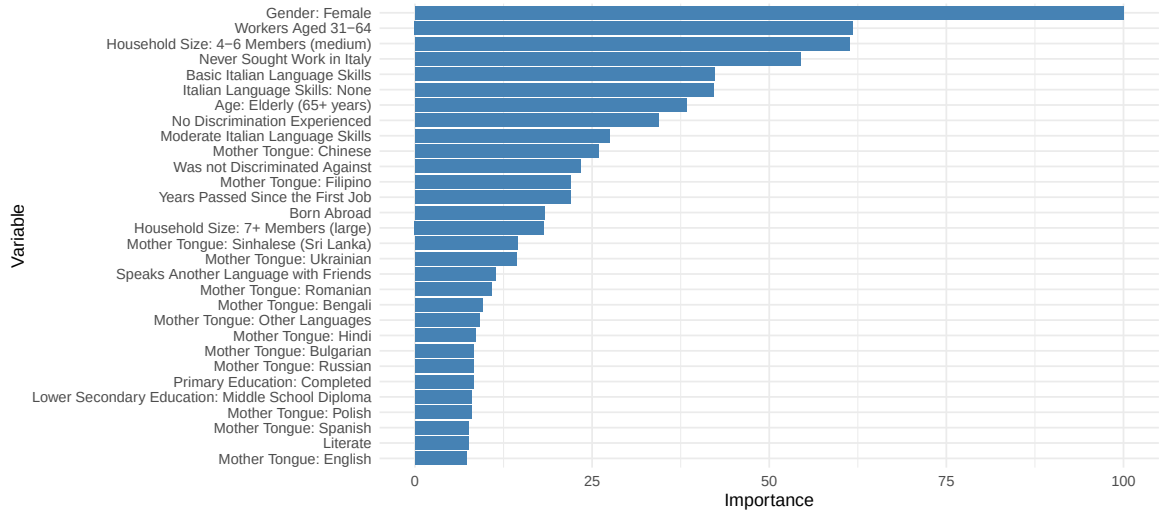

Figure 7: Variable Importance - GLM

Household size also seems to be an important factor, with both medium-sized (4-6 members) and large-sized (7+ members) households showing significant influence. This might reflect the economic pressures and resource allocation challenges faced by larger families, which can affect employment status.

As expected Italian proficiency affects employability, whereas the mother tongue appears to have a less significant impact on employment for most cases, as indicated by the relatively lower importance of various of such variables.

The GLM's sensitivity to multicollinearity, however, makes it less robust compared to models like LASSO. This sensitivity can lead to less stable coefficient estimates when predictors are highly correlated, which could be our case since we have many covariates. We decided to introduce a LASSO model to overcome overcome the issue of multicollinearity and implement a regularization penalty.

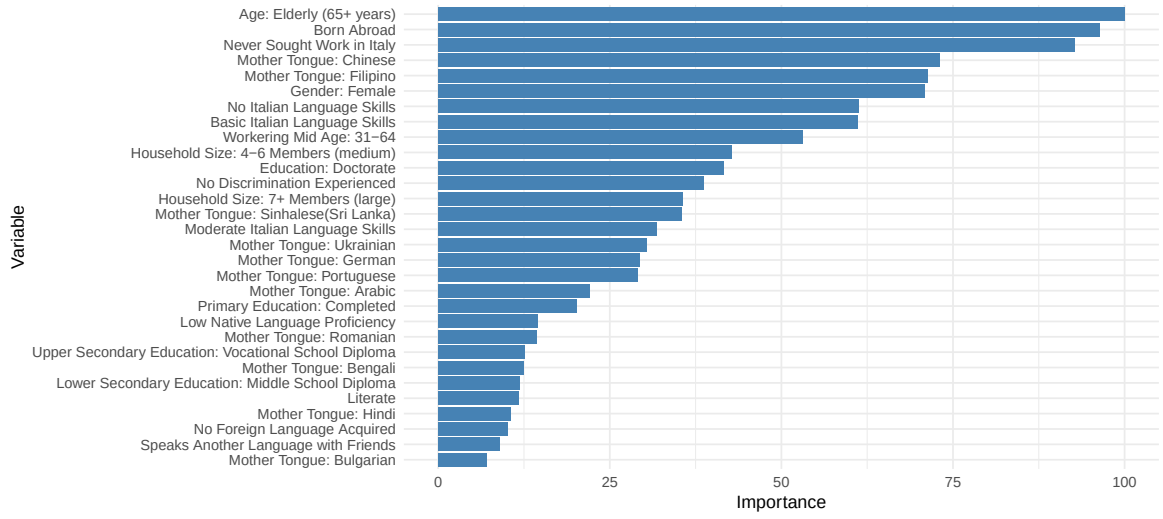

Figure 8: Variable Importance - LASSO

Figure 8 showcases the variable importance chart for the LASSO model. As mentioned, LASSO applies a penalty to prioritize the most significant variables, leading to a more refined selection. While some variables maintain similar rankings to the GLM results, there are notable differences. In the previous GLM analysis, the impact of gender appeared somewhat inflated. The LASSO model adjusts for this, giving a more balanced view. Variables like being elderly and never having sought work in Italy are highly emphasized, which makes sense as these are clear markers of unemployment. Interestingly, mother tongues such as Chinese and Filipino are highlighted as significant factors, along with different levels of Italian language proficiency. To better understand these results, examining the sign of the coefficients will reveal whether these factors positively or negatively influence employment status.

Table 1: Coefficients of the Lasso Model

Variable	Coefficient	Variable	Coefficient
Born Abroad	1.540	Speaks Another Language with Friends	0.113
Mother Tongue: Chinese	1.167	Mother Tongue: Bulgarian	0.112
Mother Tongue: Filipino	1.139	Literate	-0.168
No Italian Language Skills	0.978	Lower Secondary Education: Middle School Diploma	-0.188
Basic Italian Language Skills	0.977	Low Native Language Proficiency	-0.231
Workering Mid Age: 31-64	0.848	Primary Education: Completed	-0.231
Education: Doctorate	0.665	Mother Tongue: Arabic	-0.323
Mother Tongue: Sinhalese(Sri Lanka)	0.566	Mother Tongue: Portuguese	-0.352
Moderate Italian Language Skills	0.509	Mother Tongue: German	-0.468
Mother Tongue: Ukrainian	0.486	Household Size: 7+ Members (large)	-0.568
Mother Tongue: Romanian	0.202	No Discrimination Experienced	-0.619
Upper Secondary Education: Vocational School Diploma	0.200	Household Size: 4-6 Members (medium)	-0.684
Mother Tongue: Bengali	0.191	Gender: Female	-1.132
Mother Tongue: Hindi	0.163	Never Sought Work in Italy	-1.482
No Foreign Language Acquired	0.143	Age: Elderly (65+ years)	-1.599

The coefficient analysis indicates that even with enhanced model robustness, being female remains the most significant negative factor affecting employment (aside from previously mentioned variables). This underscores systemic gender differences in the labour market, possibly influenced by cultural norms, childcare responsibilities, and potential employer biases. Additionally, coming from medium and large-sized families continues to negatively impact employment outcomes.

Specific mother tongues, such as Chinese, Filipino, and Ukrainian, positively influence employment, suggesting these groups may possess skills or attributes valued in the labour market, or they might face less cultural or linguistic integration challenges. In contrast, speakers of Arabic and Portuguese face negative impacts, which could stem from higher levels of discrimination or cultural differences that are more challenging to bridge. These differences can also be influenced by the varying educational backgrounds and reasons for immigrating, which differ significantly across ethnic groups. Furthermore, proficiency in one's mother tongue also plays a role, as it can facilitate social integration and access to community networks, which indirectly support

employment opportunities.

Interestingly, very high proficiency in Italian does not significantly affect employment; basic and moderate levels of Italian seem sufficient. This suggests that additional proficiency does not substantially increase employability. This could be due to the nature of available jobs that do not require advanced language skills or the presence of ethnic enclaves where native languages are prevalent. Additionally, having only a primary or lower secondary education negatively impacts employment prospects. Naturally, this reflects on the role of educational attainment in securing better job opportunities and the importance of access to higher education for improving employability and career advancement.

6 Further Interaction Analysis

As mentioned numerous times, our findings revealed an interesting result: being female has the most significant negative impact on employment opportunities.

Literature on women's employment amongs immigrants[10] in the European Union, highlights that women face lower labour market participation rates, are more likely to hold part-time or temporary jobs, and experience a persistent wage gap compared to men, indicating that immigrant women face substantial barriers. Women often shoulder more household and caregiving responsibilities, which impacts their employment opportunities and career progression. Our study finds similar patterns, with household size and composition being critical factors in employment outcomes for immigrant women. While women have increasingly high levels of educational attainment, this does not always translate into equal employment opportunities or earnings compared to their male counterparts. This highlights the importance of education and skills training in improving women's employment prospects.

This led us to try to see if an interaction between ethnic background and age could explain some key differences specifically for women. For this part of the analysis, we subset our dataset to include only women. For the data preparation, the employment and mother tongue variables have been coded as before, whereas the age variable has been recoded to a categorical variablebetween 15 and 70 years of age with 10-year intervals. We then proceeded to fit a Logit model with interactions to capture multiple effects. After fitting the model, we decided that the best way to interpret this information would be to calculate the predicted probabilities of employment and and show that across different age groups and ethnic bacgkrounds.

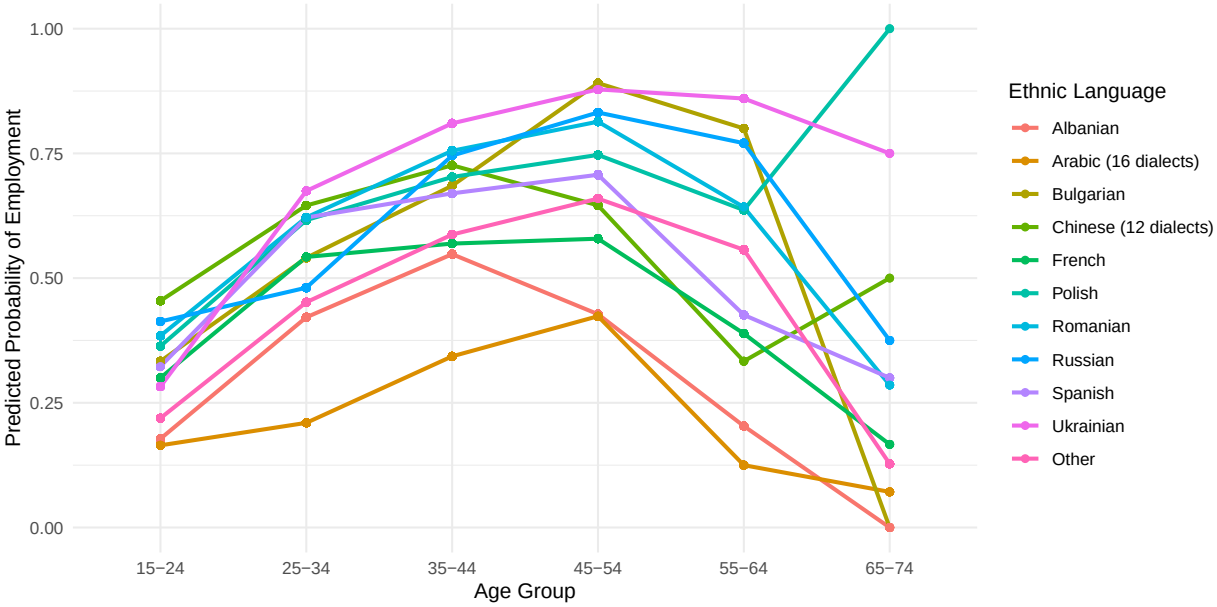

Figure 9: Interaction Effect of Mother Tongue and Age on Employment Status in Women

In Figure 9 each line represents a different ethnic language group between the ten most common ones, with the y-axis showing the predicted probability of employment and the x-axis depicting various age groups.

The EU average for employement rates in immigrant women is around 59%, while the Italian average is around 48%[11]. Italy experiences one of the worst gender disparities in Europe in terms of employment rates, with some ethnic backgrounds (Arabic speakers) of women having less than 40% employment rate across all age groups.

We can note a positive change from the "15-24" age group to the "25-34" age group, which can be explained by the initial entry into the labour market after the obtainment of tertiary education. The plot reveals that employment probability generally peaks in the "45-54" age group for most languages, indicating that late-career women have the highest employment chances. Notably, Ukrainian (significantly higher than average with peaks around 80% across three different age groups), French, Russian, Romanian, Bulgarian speakers exhibit higher employment probabilities across most age groups, suggesting these groups face fewer barriers. Conversely, Arabic and Albanian speakers show lower employment probabilities, across all the age groups, highlighting potential cultural and systemic challenges.

7 Conclusions and Findings

In our comprehensive analysis of the SCIF survey data, we explored the complex relationship between key demographic characteristics, linguistic proficiency, on employment outcomes among immigrants in Italy. Our findings indicate that language skills, educational background, and

gender significantly impact employment status, revealing various fundamental insights into the integration challenges faced by immigrants.

Proficiency in Italian turned out as a significant factor for successful integration into the labour market. The research showed that, immigrants with higher proficiency in Italian are more likely to secure employment, pointing to the importance of language integration programs. Our models provided a nuanced understanding of how these variables interact to affect employment prospects. For instance, the ability to communicate effectively in Italian is crucial, emphasizing the need for targeted language training programs that go beyond basic communication to include professional terminology and workplace-specific language skills. This approach can enhance employability and facilitate smoother integration into the workforce.

Gender disparities were particularly pronounced in our findings. Female immigrants face significant challenges in the labour market, with gender being the most substantial negative factor affecting employment opportunities. This aligns with broader European studies, which indicate that immigrant women often have lower labour market participation rates and are more likely to hold part-time or temporary jobs. Our analysis showed that household size and composition are critical factors influencing employment outcomes for immigrants, reflecting the additional responsibilities and economic pressures immigrant families face. Educational background also plays a significant role in employment outcomes. Immigrants with higher educational attainment, particularly those with secondary or tertiary education, have better employment prospects. This underscores the importance of providing access to educational opportunities and vocational training programs tailored to the needs of immigrants. These programs should aim to bridge existing skill gaps and support career advancement.

We discovered through the interaction analysis between age, gender, and ethnic background that with Italy's employment rate for immigrant women at 48%, below the EU average of 59%, certain immigrant groups face more pronounced challenges. Arabic-speaking women show the lowest employment rates, below 40% across all age groups, highlighting systemic and cultural barriers. In contrast, immigrants with European origins demonstrate higher employment probabilities, peaking in the "45-54" age group. Differences could be explained through a more specific analysis, but cultural differences are apparent. Many female immigrants encounter persistent obstacles, necessitating targeted policy interventions to enhance their labour market integration.

Based on these limited insights and current literature[12] on the topic, several policy recommendations could be suggested to enhance immigrant integration into Italy's labour market. Firstly, tailored on-the-job training programs should be implemented to bridge skill gaps and incorporate cultural sensitivity. These programs should focus on young people and women, who face the most significant barriers. Secondly, language training should be comprehensive, extending beyond basic skills to include professional and workplace-specific terminology. Thirdly, policies should address gender disparities by providing targeted support for female immigrants, such as childcare

services and flexible work arrangements. Finally, educational opportunities should be expanded, with a focus on vocational training and higher education access to improve employability and career prospects.

In conclusion, enhancing the integration of immigrants into Italy's labour market requires a challenging approach that addresses language barriers, gender disparities, and educational needs. By implementing comprehensive and culturally sensitive policies, Italy can significantly improve the economic outcomes for immigrants and their overall contribution to society. This holistic approach will not only benefit immigrants but also foster a more inclusive and cohesive society, ultimately contributing to Italy's socio-economic development.

References

1. ISTAT. (2012). *SOCIAL CONDITION AND INTEGRATION OF FOREIGN CITIZENS*. https://www.istat.it/en/archivio/191097
2. MIPEX. (2012). *Italy MIPEX indicators*. https://www.mipex.eu/italy
3. Studies, C. M. (2023). Comparative migration studies. *Springer Open*. https://comparativemigrationstudies.springeropen.com/articles/10.1186/s40878-023-00347-y
4. Andrew Geddes, P. S. (2016). *The politics of migration and immigration in europe* (Second). SAGE Publications Ltd.
5. European Union, C. of the. (2016). *Declaration of the european ministerial conference on integration (zaragoza, 15 & 16 april 2010)*. https://migrant-integration.ec.europa.eu/library-document/declaration-european-ministerial-conference-integration-zaragoza-15-16-april-2010_en
6. Conti, C. (2017). *Italy's contribution to migration statistics*. UNECE. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.10/2017/mtg1/2017_UNECE_Migration_WP_08_Italy_Conti_ENG.pdf
7. ISTAT. (2012). *Census data*. http://dati-censimentopopolazione.istat.it/Index.aspx#
8. Sydney, T. U. of. (2021). *Lasso regression*. https://bookdown.org/tpinto_home/Regularisation/lasso-regression.html
9. *Variable importance measures*. (2020). https://ema.drwhy.ai/featureImportance.html
10. OECD. (2019). *Gender differences in immigrant integration*. https://doi.org/https://doi.org/https://doi.org/10.1787/9789264307216-10-en
11. Labour, M. of, & Policies, S. (2023). *Employment of migrant workers still growing*. https://integrazionemigranti.gov.it/en-gb/Ricerca-news/Dettaglio-news/id/3341/Employment-of-migrant-workers-still-growing
12. Solano, G., & Ponzo, I. (2022). *Social policies*. Migration Policy Group. https://www.migpolgroup.com/wp-content/uploads/2022/09/Solano-Ponzo-2022-Social-Policies.pdf

A Code appendix

```
## Setup options for R Markdown

# Load libraries
library(knitr)       # provides the kable function & chunk options
library(kableExtra)  # provides kable_styling for kable settings
library(tidyverse)   # loads amongst others dplyr, tidyr and ggplot2
library(tidymodels)  # loads several packages useful for model testing
library(caret)       # variable importance estimation
library(scales)      # scale, transform, format data for visualization
library(haven)       # read and write data files used by SAS
library(readxl)      # read and write Excel files
library(tibble)      # provides a reimagining of data frames in R
library(broom)       # output of functions into tidy data frames
library(DMwR2)       # tools for handling missing values, outliers
library(openxlsx)    # reading, writing, and editing Excel file
library(forcats)     # tools for working with categorical data


# Set basic display options
options(
  digits = 3,   # limit the number of significant digits
  width  = 63   # limit the width of code output
)

# Set knitr options
opts_chunk$set(
  echo       = FALSE,    # Do not print code
  warning    = FALSE,    # Suppress warnings
  message    = FALSE,    # Suppress messages
  fig.align  = "center"  # Center figures
)

# Set a default ggplot theme
theme_set(
  theme_minimal() +
  theme(plot.title = element_text(face = "bold", size = 16,
```

```
                                    hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5, size = 14),
        axis.title = element_text(face = "bold", size = 13),
        axis.text.x = element_text(vjust = 1.5),
        legend.title = element_text(face = "bold", size = 12),
        legend.text = element_text(size = 12),
        strip.background = element_rect(fill = "black"),
        strip.text = element_text(face = "bold", size = 12,
                                  color = "white")))
# Read the dataset, it is originally in .dta format for stata

data <- read_dta("SCIF_original.dta")

# To get an idea of the variables, view all missing values in order
#(percentage - most missing to least)
NA_perc <- data %>% summarise(across(everything(), ~ mean(is.na(.)))) %>%
  pivot_longer(cols = everything(), names_to = "variable",
               values_to = "missing_pct") %>%
  arrange(desc(missing_pct))

#View(NA_perc)
kable(NA_perc)

# Since the dataset is large, keep vars with less than 60% missing val
vars_to_keep <- NA_perc %>%
  filter(missing_pct <= 0.60) %>% pull(variable)

# Subset the original data to keep only the selected vars
cleaned_data <- data[, vars_to_keep]

# Clarify which variables are categorical, since R reads all
#as numerical from .dta
cleaned_data <- cleaned_data %>%
  mutate(across(where(is.labelled), as_factor))

# View and Deal with our variable of interest which is employment
#status and recode it to 2 categories and remove missing rows
#for dependent variable
```

```
cleaned_data <- cleaned_data %>% filter(!is.na(COND5))
#summary(cleaned_data$COND5)

# Recoding using grepl to check if COND5 starts with "Occupato"
cleaned_data <- cleaned_data %>%
  mutate(COND5_recode = case_when(
    as.numeric(COND5) %in% 1:3 ~ 1, # Assuming '1', '2', '3' working
    as.numeric(COND5) %in% 4:5 ~ 0, # Assuming '4', '5' not working
    TRUE ~ 2                        # Catch-all for unexpected values
  ))

# Finally view the summary of the new variable
#table(cleaned_data$COND5_recode)

# We will perform imputation for the missing values to be able
# to fit models easier # For categorical it will be mode imputation
#and for numerical - mean imputation mode function for imputation

get_mode <- function(v) {
  uniqv <- unique(na.omit(v))
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

cleaned_data <- cleaned_data %>%
  mutate(across(where(is.numeric),
                  ~replace_na(., mean(., na.rm = TRUE))),
         across(where(is.factor), ~replace_na(.,get_mode(.))))

# Save the cleaned dataset as a new Stata file
write_dta(cleaned_data, "SCIF_cleaned.dta")

# List of variables of interest to our research question, related to
#language integration and discrimination. Which are similar
#(will portray intracorrelation ) transformed to numerical and combined
#through mean for categorization into a new vatiable
cleaned_data <- cleaned_data %>%
  mutate(across(c(INT_LI2NEW, INT_LI3NEW, INT_LI4NEW, INT_LI5NEW,
                  INT_LI6NEW, INT_LI7, INT_LI8, INT_LI9),
```

```
                ~factor(.,
                        levels = c("Molto", "Abbastanza",
                                   "Poco", "Per niente"),
                        labels = c(1, 2, 3, 4))))

cleaned_data <- cleaned_data %>%
  mutate(across(c(INT_LI2NEW, INT_LI3NEW, INT_LI4NEW, INT_LI5NEW,
                  INT_LI6NEW, INT_LI7, INT_LI8, INT_LI9), as.numeric))

cleaned_data <- cleaned_data %>%
  rowwise() %>%
  mutate(Avg_Score = round(mean(c_across(c(INT_LI2NEW, INT_LI3NEW,
  INT_LI4NEW, INT_LI5NEW, INT_LI6NEW, INT_LI7, INT_LI8, INT_LI9)),
  na.rm = TRUE), 0)) %>% ungroup()

cleaned_data <- cleaned_data %>%
  mutate(Summ_Lang_Skill = case_when(
    Avg_Score == "1" ~ "Well",
    Avg_Score == "2" ~ "Quite a bit",
    Avg_Score == "3" ~ "Not so much",
    Avg_Score == "4" ~ "Not at all" # Handles any NAs or
  ))

cleaned_data$Summ_Lang_Skill <- factor(cleaned_data$Summ_Lang_Skill,
                                levels = c("Well", "Quite a bit",
                                    "Not so much", "Not at all"))
cleaned_data2 <- cleaned_data %>%
  mutate(
    INT_LI4NEW = factor(INT_LI4NEW, levels = c(1, 2, 3, 4),
                        labels = c("Quite well", "Moderately",
                                   "Poorly", "Not at all")),
    FORM_LM1NEW = case_when(
      FORM_LM1NEW == "Albanese" ~ "Albanian",
      FORM_LM1NEW ==
        "Serbo; Croato; Bosniaco; Montenegrino" ~ "Serbo-Crotian",
      FORM_LM1NEW == "Rumeno" ~ "Romanian",
      FORM_LM1NEW == "Ucraino" ~ "Ukrainian",
      FORM_LM1NEW == "Polacco" ~ "Polish",
```

```
      FORM_LM1NEW == "Francese" ~ "French",
      FORM_LM1NEW == "Russo" ~ "Russian",
      FORM_LM1NEW == "Arabo (16 idiomi)" ~ "Arabic",
      FORM_LM1NEW == "Spagnolo" ~ "Spanish",
      FORM_LM1NEW == "Cinese (12 idiomi)" ~ "Chinese",
      TRUE ~ FORM_LM1NEW  # Default case to keep original if no match
    )
  )

cleaned_data2$INT_LI2NEWNEW <- factor(cleaned_data2$INT_LI2NEW,
                                      levels = c(1, 2, 3, 4),
                                      labels = c("A lot", "Quite",
                                                 "Little",
                                                 "Not at all"))
cleaned_data2$INT_LI3NEWNEW <- factor(cleaned_data2$INT_LI3NEW,
                                      levels = c(1, 2, 3, 4),
                                      labels = c("A lot", "Quite",
                                                 "Little",
                                                 "Not at all"))
combined_data <- rbind(
  data.frame(Skill = "Reading", Level = cleaned_data2$INT_LI2NEWNEW),
  data.frame(Skill = "Writing", Level = cleaned_data2$INT_LI3NEWNEW))

combined_data <- combined_data %>%
  group_by(Skill, Level) %>%
  summarize(Count = n()) %>%
  group_by(Skill) %>%
  mutate(Prop = Count/sum(Count))
ggplot(combined_data, aes(x = Level, y=Prop, fill = Skill,
                          label=label_percent(0.1)(Prop))) +
  geom_col(position = "stack", color = "black", na.rm = TRUE) +
  geom_text(position = position_stack(vjust = 0.5),
            color = "white", size = 3) +  # Add text labels
  labs(#title = "READING AND WRITING KNOWLEDGE OF THE ITALIAN LANGUAGE",
       x = "Level", y = "Proportion") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 40, hjust = 1)) +
  scale_fill_brewer(palette = "Set1") +
```

```r
  theme(plot.title = element_text(hjust = 0.5))
top_mother_tongues <- cleaned_data2 %>%
  group_by(FORM_LM1NEW) %>%
  summarise(total = n(), .groups = 'drop') %>%
  top_n(10, total) %>%
  pull(FORM_LM1NEW)

# Filter data for only the top 10 mother tongues
filtered_data_graph <- cleaned_data2 %>%
  filter(FORM_LM1NEW %in% top_mother_tongues)

# Count the occurrences and calculate proportions
data_to_plot <- filtered_data_graph %>%
  group_by(FORM_LM1NEW, INT_LI4NEW) %>%
  summarise(count = n()) %>%
  mutate(total = sum(count), # Calculate responses per mother tongue
         proportion = count / total)  # Calculate proportion
ggplot(data_to_plot, aes(x = FORM_LM1NEW, y = proportion,
                         fill = INT_LI4NEW)) +
  geom_bar(stat = "identity", position = "stack") +
  coord_flip() +  # Flip the coordinates to make bars horizontal
  labs(x = "Mother Tongue", y = "Proportion",
       fill = "Understanding Level",
       #title = "Level of Spoken Italian
       #(How well can an Italian understand you)"
       ) +
  theme_minimal() +
  scale_fill_brewer(palette = "Set1") +
  theme(legend.position = "top",
        plot.margin = unit(c(1, 1, 1, 1), "cm"), # Increase plot margins
        legend.box.margin = margin(0, 100, 0, 0),
        legend.margin = margin(0, 0, 0, 0),
        legend.key.size = unit(0.4, "cm"),
        legend.key.width = unit(0.4, "cm"))  # Reduce legend key size


vars_of_interest <- c("LAV_ATT27","TIT_STUD", "LAV_ATT15",
                      "LAV_ATT15BIS", "LAV_ATT22", "LAV_ATT26",
```

```
                      "LAV_ATT28", "LAV_ATT40", "LAV_ATT40BIS",
                      "ANNI_PRIMO_LAV", "COND5_recode",
                      "LAV_ITA2", "LAV_ITA4", "LAV_ITA5","DISCR_44",
                      "DISCR_47", "LAV_ATT1", "LAV_ITA1", "LAV_PER1",
                      "DISCR_44", "INT_LI1", "FORM_LM1NEW",
                      "FORM_LM1BIS", "FORM_LM2", "INT_LI9",
                      "INT_LI10",  "FORM_LM1NEW","FORM48",  "SG11",
                      "SG21", "FORM1", "FORM3", "SG4",
                      "SG13", "SG16", "INT_LI2NEW","INT_LI3NEW",
                      "INT_LI4NEW", "INT_LI5NEW", "Summ_Lang_Skill")

# Define which of these are numeric originally
numerical_vars <- c("LAV_ATT26", "LAV_ITA2", "SG21", "SG4")

# Define new dataset with vars of interest
data_interest <- cleaned_data %>%
  select(all_of(vars_of_interest))

# Transform all num variables to cat for ease of model fitting
# Starting with LAV_ATT26
# Convert to numeric, "Non sa" to NA
data_interest <- data_interest %>%
  mutate(LAV_ATT26 = as.numeric(na_if(as.character(LAV_ATT26),
                                      "Non sa")))
data_interest <- data_interest %>%
  mutate(LAV_ATT26_Cat = cut(LAV_ATT26,
                        breaks = c(1943, 1989, 2000, 2011),
                        labels = c("1944-1989", "1990-2000",
                                   "2001-2011"),
                        right = TRUE,
                        include.lowest = TRUE))
# View the changes
table(data_interest$LAV_ATT26_Cat)

# Do the same for LAV_ITA2
data_interest <- data_interest %>%
  mutate(LAV_ITA2 = round(LAV_ITA2))
data_interest <- data_interest %>%
```

```
  mutate(LAV_ITA2_Categories = cut(LAV_ITA2,
                           breaks = c(1937, 1980, 1990, 2001, 2012),
                          abels = c("1937-1980", "1981-1990",
                                    "1991-2001", "2002-2012"),
                          include.lowest = TRUE, right = FALSE))
# View the changes
table(data_interest$LAV_ITA2_Categories)
# Do the same for SG21
data_interest <- data_interest %>%
  mutate(
    SG21_Category = cut(SG21,
                        breaks = c(15,30, 64, Inf),
                        labels = c("15-30: Young Workers",
                                   "31-64: Older Workers",
                                   "65+: Elderly"),
                        right = FALSE))  # Ensures that 15 is included
# View the changes
table(data_interest$SG21_Category)

# Do the same for SG4
data_interest <- data_interest %>%
  mutate(
    SG4_Category = cut(SG4,
                       breaks = c(-Inf, 3, 6, Inf),
                       # Defining the ranges
                       labels = c("Small (up to 3 members)",
                        "Medium (4-6 members)", "Large (7 and above)"),
                       right = TRUE))
    # Ensures that 3 is included in the 'Small' category, 6 in 'Medium'
  # View the changes
table(data_interest$SG4_Category)
# clump the rest of the variables which have too many categories
data_interest <- data_interest %>%
  mutate(ANNI_PRIMO_LAV = as.numeric(ANNI_PRIMO_LAV),
         LAV_ATT27 = as.numeric(LAV_ATT27),
         LAV_ATT40 = as.numeric(LAV_ATT40),
         LAV_ATT28 = fct_lump_n(LAV_ATT28, n = 3),
         LAV_PER1 = fct_lump_n(LAV_PER1, n = 5),
```

```
        FORM_LM1NEW = fct_lump_min(FORM_LM1NEW, min = 174,
                                   other_level = "Other Mother Tongues")
  )
# Perform mutation again for redundant categories
data_interest <- data_interest %>%
  mutate(across(where(is.numeric), ~replace_na(.,
                                               mean(., na.rm = TRUE))),
         across(where(is.factor), ~replace_na(.,get_mode(.))))

vars_of_interest_cat <- c("LAV_ATT27", "LAV_ATT15", "LAV_ATT15BIS",
                          "LAV_ATT22", "LAV_ATT26_Cat", "LAV_ATT28",
                          "LAV_ATT40", "LAV_ATT40BIS", "ANNI_PRIMO_LAV",
                          "COND5_recode","LAV_ITA2_Categories",
                          "LAV_ITA4", "LAV_ITA5", "DISCR_44", "DISCR_47",
                          "LAV_ATT1", "LAV_ITA1", "LAV_PER1", "DISCR_44",
                          "INT_LI1", "FORM_LM1NEW", "FORM_LM1BIS",
                          "FORM_LM2", "INT_LI10", "FORM48", "SG11",
                          "SG21_Category", "FORM1", "FORM3",
                          "SG4_Category","SG13", "SG16",
                          "Summ_Lang_Skill", "TIT_STUD")

# From the interesting variables remove all the ones related to work,
# since they will be highly correlated and won't make for good covariates
data_interest_cat <- data_interest %>%
  select(all_of(vars_of_interest_cat)) %>%
  select(-all_of(starts_with("LAV")))


# GLM
mod_glm <- train(as.factor(COND5_recode) ~ .,
                 data = data_interest_cat,
                 method = "glm", family = "binomial")
mod_glm
view(varImp(mod_glm, scale = T))
summary(mod_glm)

renaming_dict_glm <- list(
  "SG11Femmina" = "Gender: Female",
```

```
  "SG21_Category31-64: Older Workers" = "Workers Aged 31-64",
  "SG4_CategoryMedium (4-6 members)" =
    "Household Size: 4-6 Members (medium)",
  "DISCR_44In Italia non ho mai cercato lavoro" =
    "Never Sought Work in Italy",
  "Summ_Lang_SkillNot so much" = "Basic Italian Language Skills",
  "Summ_Lang_SkillNot at all" = "Italian Language Skills: None",
  "SG21_Category65+: Elderly" = "Age: Elderly (65+ years)",
  "DISCR_47No" = "No Discrimination Experienced",
  "Summ_Lang_SkillQuite a bit" = "Moderate Italian Language Skills",
  "FORM_LM1NEWCinese (12 idiomi)" = "Mother Tongue: Chinese",
  "DISCR_44No" = "Was not Discriminated Against",
  "FORM_LM1NEWFilippino" = "Mother Tongue: Filipino",
  "ANNI_PRIMO_LAV" = "Years Passed Since the First Job",
  "SG13All estero" = "Born Abroad",
  "SG4_CategoryLarge (7 and above)" =
    "Household Size: 7+ Members (large)",
  "FORM_LM1NEWSingalese" = "Mother Tongue: Sinhalese (Sri Lanka)",
  "FORM_LM1NEWUcraino" = "Mother Tongue: Ukrainian",
  "INT_LI10Altra lingua" = "Speaks Another Language with Friends",
  "FORM_LM1NEWRumeno" = "Mother Tongue: Romanian",
  "FORM_LM1NEWBengalese" = "Mother Tongue: Bengali",
  "FORM_LM1NEWOther Mother Tongues" = "Mother Tongue: Other Languages",
  "FORM_LM1NEWHindi" = "Mother Tongue: Hindi",
  "FORM_LM1NEWBulgaro" = "Mother Tongue: Bulgarian",
  "FORM_LM1NEWRusso" = "Mother Tongue: Russian",
  "TIT_STUDIstruzione primaria (Licenza elementare)" =
    "Primary Education: Completed",
  "TIT_STUDIstruzione secondaria inferiore (Licenza media)" =
    "Lower Secondary Education: Middle School Diploma",
  "FORM_LM1NEWPolacco" = "Mother Tongue: Polish",
  "FORM_LM1NEWSpagnolo" = "Mother Tongue: Spanish",
  "FORM1No, ma sa leggere e scrivere" = "Literate",
  "FORM_LM1NEWInglese" = "Mother Tongue: English")
#,"FORM_LM1BISNo" = "Low Native Language Proficiency")

imp_glm <- varImp(mod_glm, scale = TRUE)
final_glm <- mod_glm$finalModel
```

```r
# Linear Models: the absolute value of the t-statistic for each
# model parameter is used for var importance

# Define the dataframe of importance, and again cut at top 30
imp_glm_df <- as.data.frame(imp_glm$importance) %>%
  rownames_to_column(var = "Variable") %>%
  arrange(desc(Overall)) %>%
  slice(1:30) %>%
  mutate(Variable = factor(Variable, levels = Variable))

# Rename variables using the dictionary
imp_glm_df$ReadVar <- sapply(imp_glm_df$Variable,
                             function(x) renaming_dict_glm[[x]])
kable(imp_glm_df, caption = "Top 30 Important Variables from GLM Model")

coefficients_glm <- final_glm$coefficients
keys_to_filter <- names(renaming_dict_glm)
imp_glm_df$Coefficient <-
  coefficients_glm[as.character(imp_glm_df$Variable)]
ggplot(imp_glm_df, aes(x = reorder(ReadVar, Overall), y = Overall)) +
  geom_col(fill = "steelblue") +
  theme_minimal() +
  labs(#title = "Variable Importance - GLM",
    x = "Variable",
       y = "Importance") +
  coord_flip()  # Flips the axes to make labels more readable
imp_glm_df %>%
  select(ReadVar, Coefficient, Overall) %>%
  arrange(desc(Coefficient)) %>%
  select(-Overall) %>%
  kable(caption = "Coefficients of the GLM Model",
        col.names = c("Variable", "Coefficient")) %>%
  kable_styling(latex_options = c("striped","HOLD_position",
                                  "scale_down"),
                stripe_color = "gray!30")
# LASSO
# The absolute value of the coefficients corresponding
#the the tuned model are used
```

```
# Define a renaming dictionary for Lasso for the table output
mod_lasso <- train(as.factor(COND5_recode) ~ .,
                   data = data_interest_cat,
                   method = "glmnet")
mod_lasso
view(varImp(mod_lasso, scale=T))
summary(mod_lasso)
renaming_dict_LASSO <- list(
  "SG21_Category65+: Elderly" = "Age: Elderly (65+ years)",
  "SG13All estero" = "Born Abroad",
  "DISCR_44In Italia non ho mai cercato lavoro" =
    "Never Sought Work in Italy",
  "FORM_LM1NEWCinese (12 idiomi)" = "Mother Tongue: Chinese",
  "FORM_LM1NEWFilippino" = "Mother Tongue: Filipino",
  "SG11Femmina" = "Gender: Female",
  "Summ_Lang_SkillNot at all" = "No Italian Language Skills",
  "Summ_Lang_SkillNot so much" = "Basic Italian Language Skills",
  "SG21_Category31-64: Older Workers" = "Workering Mid Age: 31-64",
  "SG4_CategoryMedium (4-6 members)" =
    "Household Size: 4-6 Members (medium)",
  "TIT_STUDDottorato di ricerca" = "Education: Doctorate",
  "DISCR_47No" = "No Discrimination Experienced",
  "SG4_CategoryLarge (7 and above)" =
    "Household Size: 7+ Members (large)",
  "FORM_LM1NEWSingalese" = "Mother Tongue: Sinhalese(Sri Lanka)",
  "Summ_Lang_SkillQuite a bit" = "Moderate Italian Language Skills",
  "FORM_LM1NEWUcraino" = "Mother Tongue: Ukrainian",
  "FORM_LM1NEWTedesco" = "Mother Tongue: German",
  "FORM_LM1NEWPortoghese" = "Mother Tongue: Portuguese",
  "FORM_LM1NEWArabo (16 idiomi)" = "Mother Tongue: Arabic",
  "TIT_STUDIstruzione primaria (Licenza elementare)" =
    "Primary Education: Completed",
  "FORM_LM1BISNo" = "Low Native Language Proficiency",
  "FORM_LM1NEWRumeno" = "Mother Tongue: Romanian",
  "TIT_STUDIstruzione professionale (scuole professionali)" =
    "Upper Secondary Education: Vocational School Diploma",
  "FORM_LM1NEWBengalese" = "Mother Tongue: Bengali",
  "TIT_STUDIstruzione secondaria inferiore (Licenza media)" =
```

```
    "Lower Secondary Education: Middle School Diploma" ,
  "FORM1No, ma sa leggere e scrivere" = "Literate",
  "FORM_LM1NEWHindi" = "Mother Tongue: Hindi",
  "FORM48No" = "No Foreign Language Acquired",
  "INT_LI10Altra lingua" = "Speaks Another Language with Friends",
  "FORM_LM1NEWBulgaro" = "Mother Tongue: Bulgarian")

# Store the variable importance factor
imp_lasso <- varImp(mod_lasso, scale = TRUE)

# Make a dataframe, cutting the top 30 variables
imp_lasso
imp_df_lasso <- as.data.frame(imp_lasso$importance) %>%
  rownames_to_column(var = "Variable") %>%
  arrange(desc(Overall)) %>%
  slice(1:30) %>%
  mutate(Variable = factor(Variable, levels = Variable))

imp_df_lasso$Variable

# Rename variables using the dictionary
imp_df_lasso$ReadableVariable <- sapply(imp_df_lasso$Variable,
                                function(x) renaming_dict_LASSO[[x]])

# Convert the variable importance to a data frame
imp_df_lasso <- imp_df_lasso %>%
  arrange(desc(Overall)) %>%
  mutate(ReadableVariable = factor(ReadableVariable,
                                   levels = unique(ReadableVariable)))
ggplot(imp_df_lasso, aes(x = reorder(ReadableVariable, Overall),
                         y = Overall)) +
  geom_col(fill = "steelblue") +
  theme_minimal() +
  labs(#title = "Variable Importance - LASSO",
    x = "Variable",
       y = "Importance") +
  coord_flip()  # Flips the axes to make labels more readable
final_lasso <- mod_lasso$finalModel
```

```r
optimal_lambda <- mod_lasso$bestTune$lambda
coefficients_lasso <- coef(final_lasso, s = optimal_lambda)

# Convert to a regular dense matrix
coefficients_matrix_lasso <- as.matrix(coefficients_lasso)

# Create a data frame from the matrix
coefficients_df_lasso <- data.frame(
  Variable = rownames(coefficients_matrix_lasso),
  Coefficient = coefficients_matrix_lasso[, 1]
)

# Define the coefficients to filter through
technical_names_lasso <- names(renaming_dict_LASSO)
filtered_lasso_df <-
  coefficients_df_lasso[rownames(coefficients_df_lasso) %in%
                          technical_names_lasso, , drop = FALSE]
filtered_lasso_df <-
  filtered_lasso_df[order(-filtered_lasso_df$Coefficient), ]
filtered_lasso_df$ReadableName <-
  renaming_dict_LASSO[rownames(filtered_lasso_df)]

# Filter dataframe to only include those names and show it in a table

filtered_lasso_df_norows <- filtered_lasso_df
rownames(filtered_lasso_df_norows) <- NULL
kable(list(filtered_lasso_df_norows[1:15,
                    c("ReadableName", "Coefficient")],
            filtered_lasso_df_norows[16:30,
                   c("ReadableName", "Coefficient")]),
      caption = "Coefficients of the Lasso Model",
      col.names = c("Variable", "Coefficient"),
      row.names = FALSE,
      booktabs = TRUE, valign = 't') %>%
  kable_styling(latex_options = c("striped","HOLD_position"),
                font_size = 7)

# Understanding interaction between gender and age group
```

```r
# for only the subset of women
women_data <- cleaned_data %>%
  filter(SG11 == "Femmina")

women_lang_map <- c(
  "Albanese" = "Albanian",
  "Arabo (16 idiomi)" = "Arabic (16 dialects)",
  "Bulgaro" = "Bulgarian",
  "Cinese (12 idiomi)" = "Chinese (12 dialects)",
  "Francese" = "French",
  "Polacco" = "Polish",
  "Rumeno" = "Romanian",
  "Russo" = "Russian",
  "Spagnolo" = "Spanish",
  "Ucraino" = "Ukrainian",
  "Other" = "Other"
)

women_data2 <- women_data %>%
  filter(SG21 >= 15, SG21 <= 70) %>%
  mutate(age_group = cut(SG21,
                         breaks = seq(15, 75, by = 10),
                         include.lowest = TRUE,
                         labels = paste(seq(15, 65, by = 10),
                                        seq(24, 74, by = 10),
                                        sep = "-"),
                         right = TRUE))

# Recode languages to top 10 languages, excluding "Italiano"
women_data2 <- women_data2 %>%
  mutate(FORM_LM1NEWo = ifelse(FORM_LM1NEW == "Italiano", "Other",
                               as.character(FORM_LM1NEW)),
         FORM_LM1NEW_w = fct_lump_n(as.factor(FORM_LM1NEWo), n = 10,
                                    other_level = "Other"))

women_data2 <- women_data2 %>%
  mutate(FORM_LM1NEW_w = recode(FORM_LM1NEW_w, !!!women_lang_map))
```

```r
# Convert necessary columns to factors
women_data2 <- women_data2 %>%
  mutate(FORM_LM1NEW_w = as.factor(FORM_LM1NEW_w),
         age_group = as.factor(age_group),
         COND5_recode = as.factor(COND5_recode))

#  Fit logistic regression model with interaction term
women_model <- glm(COND5_recode ~ FORM_LM1NEW_w * age_group,
                   family = binomial(link = "logit"),
                   data = women_data2)

#  Summarize the model
summary(women_model)

#  Predict probabilities
women_data2$predicted_prob <- predict(women_model, type = "response")
ggplot(women_data2, aes(x = age_group, y = predicted_prob,
                        color = FORM_LM1NEW_w,
                        group = FORM_LM1NEW_w)) +
  geom_line(linewidth = 1) +
  geom_point() +
  labs(#title = "Interaction Effect of Ethnic Language and
       #Age Group on Employment Status",
       x = "Age Group",
       y = "Predicted Probability of Employment",
       color = "Ethnic Language") +
  theme_minimal() +
  theme(legend.position = "right")
```