Worldwide Labour Migration Analysis using LinkedIn Data

Team Members: Sudhish Subramaniam, Urvashi Dube, Dishant Rajesh Gupta, Dhaval Jariwala, Yesha Dharmesh Gosaliya

Overview

In today's world, people migrate from one place to the other for growth and development. The most common purpose of migration is a money hike. According to statistics, not every person who migrates has profited. When people leave their homeland they feel stuck, lonely, and depressed, and face problems such as lack of opportunities, coping with a different common language, homesickness, lack of family support, and cultural differences in the migrated country. All these problems lead to underperformance in their career and ultimately not completing their goal of a money hike.

Dataset link: https://datacatalog.worldbank.org/search/dataset/0038044/Talent-Migration---LinkedIn-Data-.The dataset is extracted from LinkedIn's Economic Graph Initiative and World Bank Group, and is available at country, country-industry and country skill level based migrations. The migration in the dataset is recorded when a LinkedIn member changes location on their LinkedIn profile. All three datasets Country migration, industry migration and skill migration are extracted from the LinkedIn profiles who have migrated based on country, industry and skills. All three datasets will be used for extensive exploratory data analysis. The analysis and predictions from the project will help people who plan to migrate to a country and also companies in the target country to analyse people coming from the base country. The analysis can also be used to examine the economy of countries. Keeping this in mind, the project aims to analyse the netflow of people in the base country and target country for five consecutive years from 2015 to 2019 and predict the netflow for the year 2023.

Methods

On extracting the data, as a first step, we will drop the unimportant columns for the project, skill_group_id from skill migration, isic_section_index, isic_section_name and industry_id from industry migration. Now, we will be managing the null values. Once the null values are removed or replaced according to requirement, the dataset is ready for analysis. Using the three datasets, country migration, industry migration and skill migration we will do the following exploratory data analysis on RShiny by creating a world map and plotting, worldwide labour migration based on:

- base country (among 140 countries) , target country (among 140 countries) and regions
- industry (143 industries)
- income (4 categories of income)
- skills (5 skill group categories and 249 skill group names)

All three datasets will be used for visualization. As per requirements datasets will be joined either by using left join, right join, inner join or full join functions of the R language. The column net_per_10k_YYYY describes the netflow_YYYY divided by total_member_ct_YYYY concerning base_country, country_name and industry_name, country_name and skill_group_name for country, industry and skill migration datasets respectively. Using this data from 2015-2019, a machine learning model will be implemented to predict the netflow for each particular country, industry and skill for the year 2023. The dataset contains data only till the year 2019, we will be using this as an advantage. Due to the pandemic in the year 2020, labour migration was completely decreased and disrupted till 2022. The data for the years 2020, 2021 and 2022 will not be apt to predict the netflow for the coming years. Hence we use the data from 2015-2019 to predict the netflow for the upcoming year, 2023. On analyzing the pattern of the data, machine learning models such as linear regression, support vector regression, random forest regression, etc will be applied to predict the netflow. All the results from the model and plots will be visualized on RShiny using line plot (plot), bar plot, scatter plots, boxplots, etc. Multiple models with various variables will be performed, compared, and assessed for each of the aforementioned models using accuracy measures including RMSE (linear regression), accuracy.