PPOL 670 Project

Female Labor Force Participation

Liping Wang

12/14/2019

Contents

Introduction

In almost every country in the world, women are less likely to participate in the labor market than men. There are various push and pull factors that explain why women work less than men. Taking care of children and family members, doing household chores, and all other activities that are outside of the formal labor market can prevent women from pursuing employment opportunities. Social and cultural norms that discourage female's labor force

participation are also prevalent in some countries. Women are more likely to work if the legal and policy frameworks are supportive of women's labor market engagement, such as flexibility of working hours and environment, taxation, and family support services.

Higher levels of female labor force participation are expected to fuel household income and overall economic growth because of more workers contributing to the economy. Women's economic empowerment may bring them more bargaining power and autonomy, especially in developing countries (Anderson et al., 2009). As labor force participation directly affects a country's economic growth, the benefits of increasing women's participation in labor markets are expected to be phenomenal for the overall economy. Theoretically, the relationship between female labor force participation and GDP per capita is hypothesized to be U-shaped (2019). Among the low-income countries, female labor force participation rates (LFPR) are the highest because they are often engaged in labor-intensive agricultural activities. As GDP rises, women whose households experience income growth might prefer activities outside of the labor market, causing the female LFPR to decline. Furthermore, once economic development and industrialization shift more jobs from farms to factories, the female LFPR will start to climb. This positive relationship between the female LFPR and GDP per capita often occurs among middle-income to high-income countries because of rapid economic growth, increased female education, and decreased fertility rates (Klasen, 2019).

While women's labor force participation has risen in many countries, rates remain quite low in some upper-middle-income countries and regions where the trends are expected to grow. Certainly, GDP is not the only factor that determines female labor force participation. Therefore, this project aims to explore what are some of the important factors that explain a country's female labor force participation. Understanding these determinants of female labor force participation will help policymakers to remove potential barriers for women, which are important for economic growth. My goal is to build a model that can explain the majority of the variation in female LFPR using the data before 2015. The success of the project will be determined by the model accuracy using the test data between 2015 and 2018.

I started this project by thinking about the factors that could have a potential influence on the level of labor force participation for females in a country. Then I collected data on those indicators to build my own dataset, including merging data from different sources. After data wrangling and cleaning, I explored the relationships among some explanatory variables and examined variable importance through visualization. Using the training dataset, I tried different types of supervised machine learning algorithms, such as the k-nearest neighbors and random forest, to determine the best algorithm for my model. In the last section, I discussed my findings, the overall project success, and recommendation on the next steps.

Problem Statement and Background

As stated above, the goal of this project is to explore the determinants of female labor force participation and build a model that can explain the majority of the variation in female LFPR.

A considerable amount of literature provides empirical evidence on the determinants of female labor force participation in the context of a specific country. Studies hardly focus on cross-country analysis, and many only investigate the relationship between one broad factor and female labor force particiaption, such as GDP and religion. However, Mehmood et al. (2015) develop a generalized model for the factors that affect female labor force participation in Muslim countries. Their results show that education attinment, especially tertiary education, has a positive link with female LFPR. They also find that as the number of children in the family increases, the female LDPR decreases, which is not surprising. Interestingly, they discover a positive relationship between inflation and female participation in the labor market. One possible explanation is that the increasing cost of living driven by inflation puts financial pressure on females and pushes them to bring more income to their households. Besides, Bayanpourtehran and Sylwester (2012) conduct a cross-country analysis to examine whether female LFPR is dependent on the religion practiced in these

countries. They conclude that countries where Protestantism is prevalent or where no religion is practiced have higher female LFPR, but the relationship between female LFPR and religion has weakened over time.

Data

The majority of the data in this project comes from the built-in "wbstats" R-package that contains World Development Indicators collected by the World Bank Group. The detailed list of indicators selected for this project is shown below.

Variable	Description
gdp	GDP per capita (current US$)
inflation	Inflation, consumer prices (annual %)
gpi	Global Peace Index (GPI)
religion	A vector of religion variables
male_unemploy	Unemployment, male (% of male labor force)
literacy	Literacy rate, adult female (% of females ages 15 and above)
fertility	Fertility rate, total (births per woman)
housework	% of time spent on unpaid domestic and care work, female
fam_plan	Contraceptive prevalence, any methods (% of women ages 15-49)
compulsory_educ	Compulsory education, duration (years)
educ_exp	Total % of Government expenditure on education
primary_enroll	School enrollment, primary, female (% net)
secondary_enroll	School enrollment, secondary, female (% net)
tertiary_enroll	School enrollment, tertiary, female (% gross)

Besides, I collected a vector of variables that measure the religious composition by country in 2010, including Buddhists, Christians, Folk Religions, Hindus, Jews, Muslims, Other

and Unaffiliated from Pew Research Center. I also scraped the Global Peace Index from Wikipedia, which measures the relative position of each nation and region's peacefulness. Since GPI assesses the level of safety and security in society, it is a good indicator to capture women's difficulties in commuting to work in a country.

The unit of analysis for this project is country-year, and the main variable of interest is the female labor force participation rate (% of female population ages 15+). Before merging the data from three different sources, I ensured the unit of analysis in each dataset is country-year. Additional data cleaning needed to be done in the religion dataset, where the values for each variable contain "<" or ">" and are non-numeric. For the GPI dataset, I removed all irrelevant information and 2019 data and transformed the wide-format data into the long format. Lastly, I combined all three datasets using full-join to create the master dataset that includes data between 2000 and 2018 for this project.

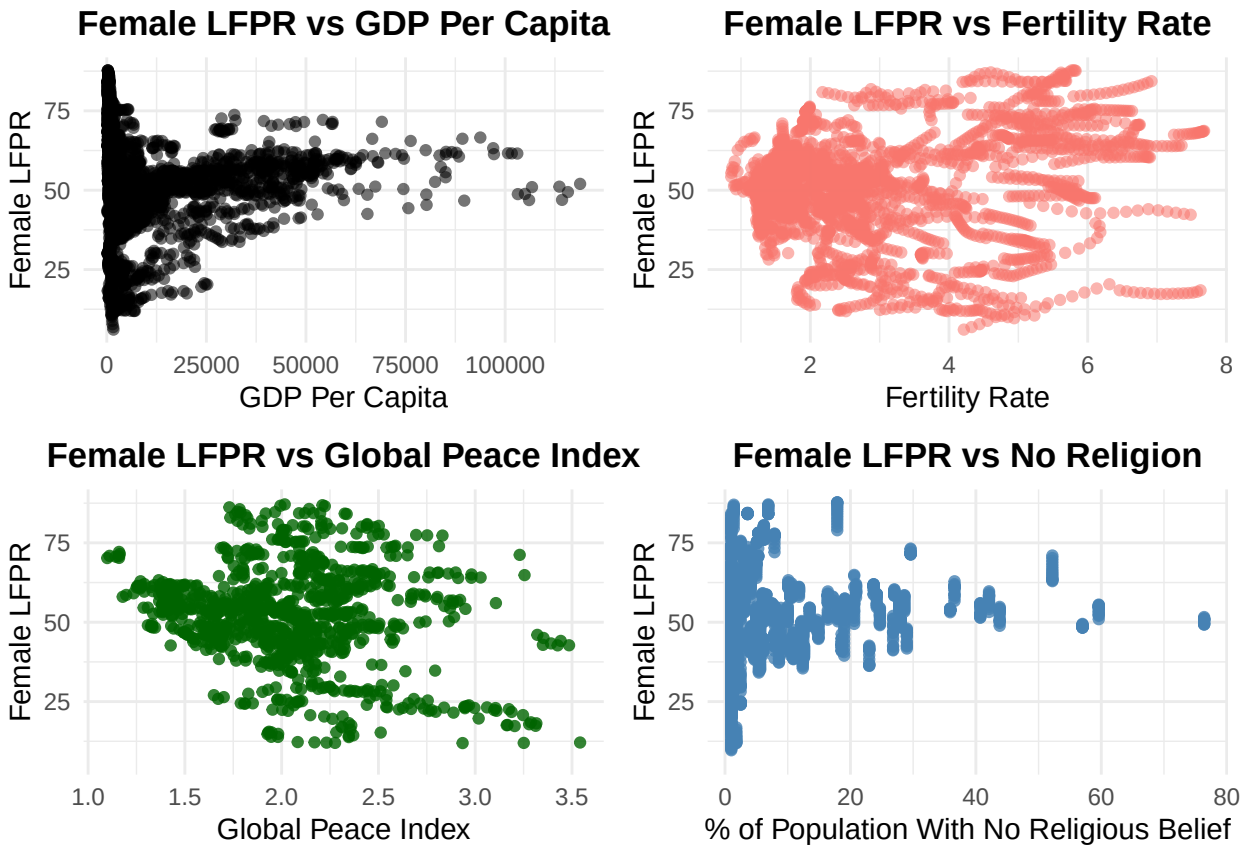
The description of each variable is presented in the above table. It is worth noting that the net enrollment rate is the ratio of children of official school age who are enrolled in school to the total population of the corresponding official school age. I chose the net rate for primary and secondary education because it is more accurate in terms of capturing the individual country's coverage and internal efficiency of each level of the education system. I selected the gross rate for tertiary education because it requires the completion of education at the secondary level and often can be pursued without age restriction. For the Global Peace Index, nations are considered more peaceful if they have lower index scores.

Due to the nature of this dataset, many variables have missing values because they are from surveys that are only conducted once in several years. Variables that contain missing values should not be dropped in this case since the non-missing values may provide valuable information to my analysis. Therefore, for all the variables related to religion, which are only from 2010, I filled in missing values with the same value from 2010 for each country. That means if the United States had 80% of Christians in 2010, then I assumed it had the same percentage of Christians in all other years since religion composition in each country does

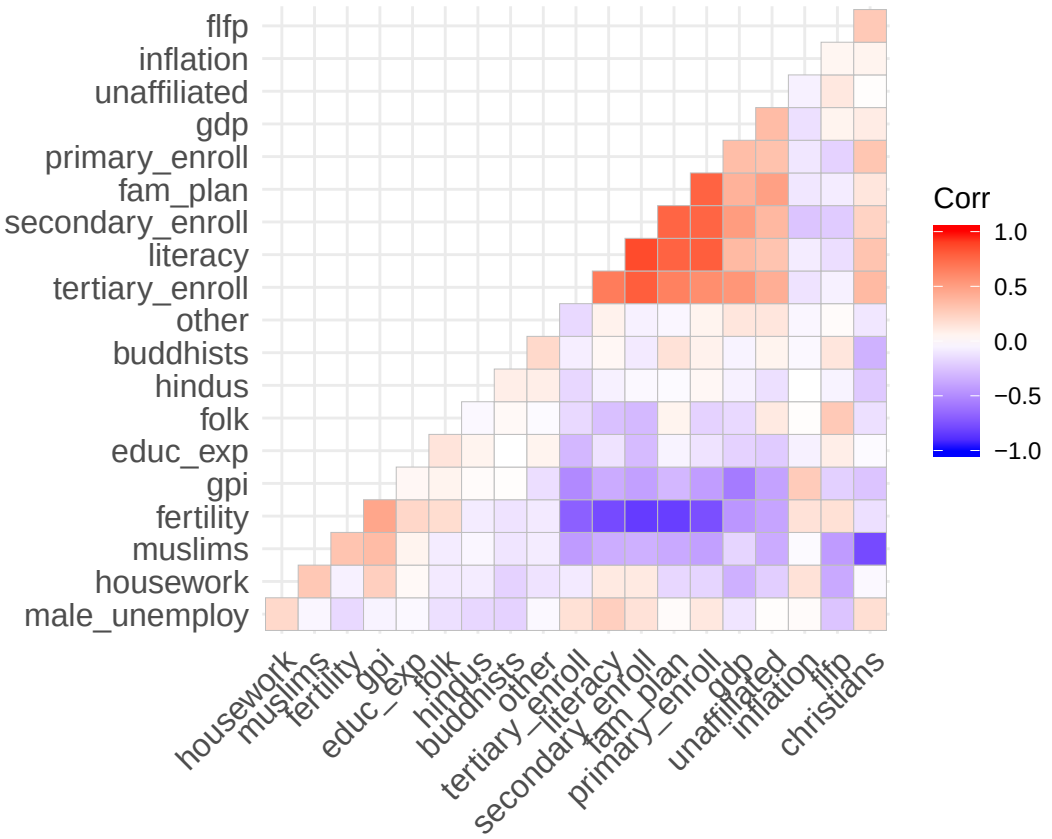
not vary much over time and can be very different across countries. For all other missing values, I used the K nearest neighbor algorithm to impute missing data by finding the k closest neighbors to the observation with missing data and then imputing them based on the non-missing values in the neighbors. The reason why I did not impute those missing values by using the information from the same country was that some countries were never surveyed to collect information about variables like housework and primary school enrollment from 2000 to 2018. However, variables like housework and primary school enrollment can be estimated based on other variables in the dataset, such as GDP per capita, years of compulsory education, literacy rate, etc.

Analysis

Before I started my analysis, I first split my master dataset into training data (before 2015) and test data (after 2014). Then I examined the distribution of my dependent variable - flfp and other selected variables. Both of the dependent variable and independent variables have lots of (good) variations. I also explored the relationships between the female LFPR and some other independent variables.



The graph that shows the correlation between the indicators is presented below. Next, I tried different algorithms to build a machine learning model to see how well it can predict the female LFPR. The methods I explored were multiple linear regression model, k-nearest neighbors, and random forest.



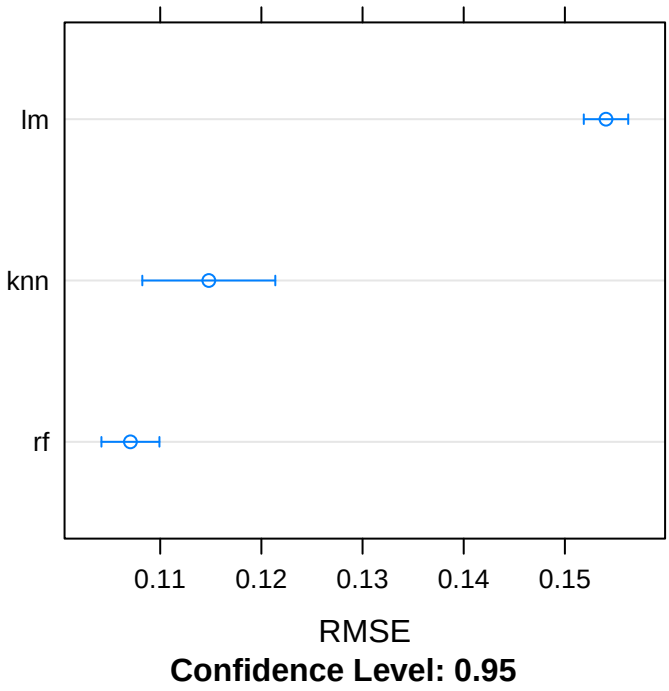
The multiple linear regression model assumes the form of f(x) is linear, which means the relationship between female LFPR and the explanatory (independent) variables is linear. The linear relationship is a strong assumption and probably will not hold since there are many outliers in the dataset, as shown in the previous graphs. However, this model is straightforward and easy to interpret. The k-nearest neighbors model works by searching through the entire training dataset for the K closet neighbors and summarizing the output variable for those K neighbors. This algorithm is easy to implement and requires no training before making predictions. Therefore, adding new data will not impact the accuracy of this method. The idea behind the random forest model is to combine many decision trees into a single model, which helps me to improve my predictions by gathering information from each decision tree. Decision tree is a method for classifying subjects into groups. It will work well if the female LFPR can be clustered into groups based on the other variables in the dataset. All three models have advantages and disadvantages, so I tried all three to see which one

has better predictive performance.

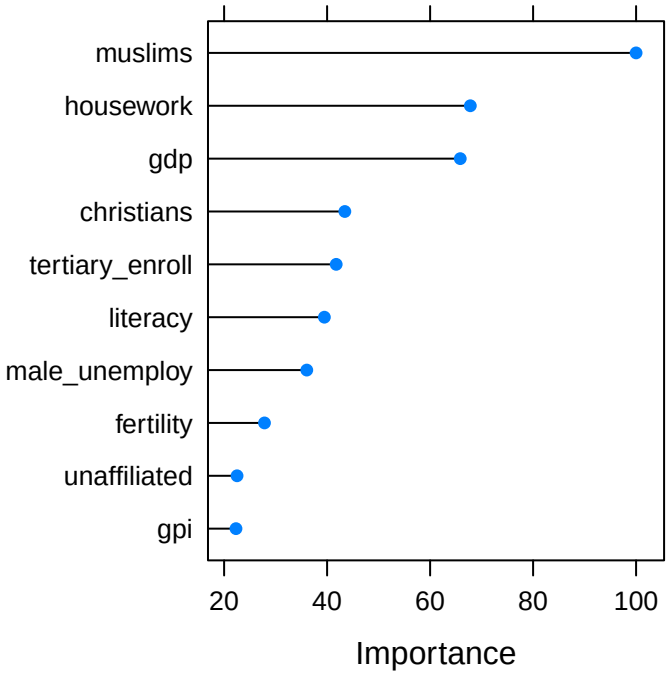
Results

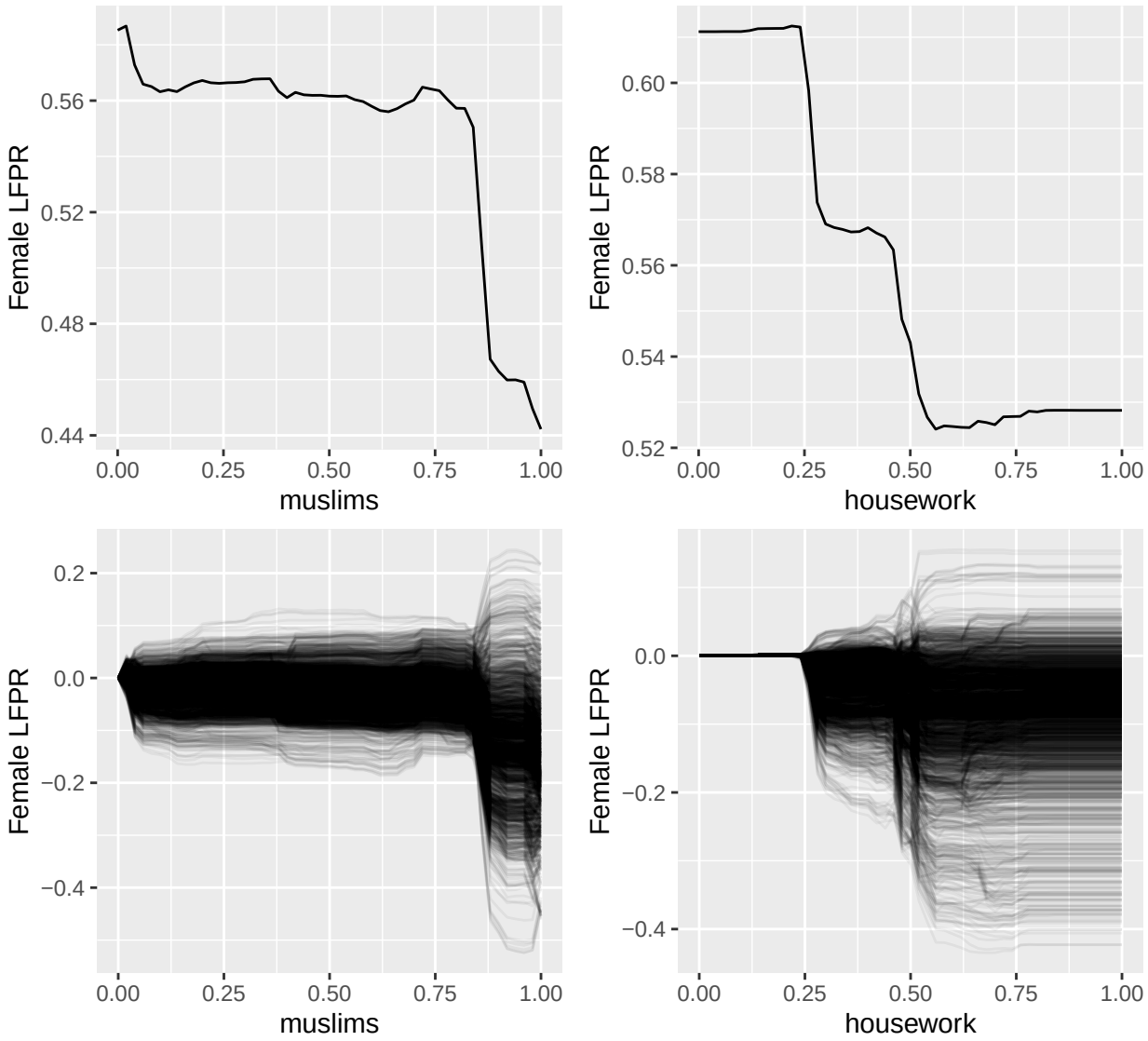
After running all three models, the Random Forest algorithm has the best performance in terms of predicting the female LFPR, which has the smallest RMSE. The RMSE is the standard deviation of the residuals, which measures the prediction errors. The RMSE in the Random Forest model is 0.107, while the RMSE for the other two models is over 0.11. The Linear Regression model produced the largest RMSE, which is the result as expected. The R-squared in the Random Forest model is 0.72, which means 72% of the variation in female LFPR can be explained by the existing independent variables in the model.

Then I used the data from 2015 to 2018 to test the Random Forest model's prediction accuracy. The result using the test data shows that the RMSE is 0.0997, which is lower than the value from training data. The similarity in the two RMSE values indicates that the Random Forest model does well in estimating the female LFPR.



For the next step, I examined the variable importance. I found that the five most important variables that can explain the female LFPR are Muslims, housework, gdp, Christians, and the tertiary enrollment rate for females. The results indicate that religion is an important factor that affects a country's female labor force participtiion. By further examining the marginal effect of the percentage of Muslims in a country on female LFPR, I found that when the percentage of Muslims is greater than 75%, female LFPR starts to decrease dramatically. This result explains why the female labor force participation in Turkey is exceptionally low compared to international standards. For the second important variable "housework", the marginal effect on female LFPR is similar. When women spend a significant amount of time on household chores, their labor force participation decreases. However, it seems that there is heterogeneity in the prediction, as shown in the two graphs below. Reasons for heterogeneity need to be further examined.





Discussion

As I stated at the beginning of this report, the success of the project is determined by whether I can find a model that explains the most variations in female LFPR and high model accuracy. Since the Random Forest model explained 72% of the variations in female LFPR, the model is considered as a success. And the low RMSE value using the test data showed the model accuracy is high. Among all the methods we learned in class, I did not run the Regression Trees model because it is similar to the Random Forest model, and based on the rule of thumb, the Random Forest model usually performs better.

In this project, I did not try different ways to impute missing values other than using the k-nearest neighbor method, which may not be the best practice to deal with the missing values. Besides, since the Random Forest model only explains the 72% variations in female LFPR, the left 28% variations could be due to other factors that I did not collect, such as indicators related to the legal framework in the country. Therefore, I could expand my analysis to getting more data on potential factors that can affect female labor force particiation.

Reference

Anderson, Siwan & Eswaran, Mukesh, 2009. "What determines female autonomy? Evidence from Bangladesh," Journal of Development Economics, Elsevier, vol. 90(2), pages 179-191, November.

Bayanpourtehrani, G., & Sylwester, K. (2013). Female labour force participation and religion: A cross-country analysis. Bulletin of Economic Research, 65(2), 107-133. doi: 10.1111/j.1467-8586.2012.00443.x

Council of Economic Advisers. (2019). Relationship between female labor force participation rates and GDP. Retrieved from https://www.whitehouse.gov/articles/relationship-female-labor-force-participation-rates-gdp/

Klasen, S. (2019). What explains uneven female labor force participation levels and trends in developing countries? The World Bank Research Observer, 34(2), 161-197. doi: 10.1093/wbro/lkz005

Mehmood, B. (2015). What derives female labor force participation in muslim countries? Retrieved from http://www.econis.eu/PPNSET?PPN=1040633110

Pew Research Center Data:https://www.pewforum.org/2015/04/02/religious-projection-table/2010/percent/all/

Wikipedia Global Peace Index Data: https://en.wikipedia.org/wiki/Global_Peace_Index

citation(tidyverse)

citation(dplyr)

citation(ggplot2)

citation(wbstats)

citation(rvest)

citation(readxl)

citation(stringr)

citation(tibble)

citation(stargazer)

citation(gridExtra)

```
citation(recipes)
citation(caret)
citation(vip)
citation(pdp)
```