Predicting an Innovative Firm - (CYO) Project

Myint Moe Chit

6/16/2019

Introduction

Innovation is considered as one of the essential components of a firm's competitiveness. Innovation of businesses is also important for economies. An innovative firm is likely to generate more profits, create more skilled-jobs, and pay higher wages. It is importance for investors, managers and policy makers to be able to identify an innovative firm. In this project, innovation of a firm is measured in a broader sense. A firm is considered innovative if it can introduce a new product or service.

In this project, I apply machine learning techniques to predict innovative firms using firm-level data from the World Bank's Enterprise Surveys (WBES). The surveys were conducted in various countries between 2007 and 2018 (Data release date: May 6, 2019). These cross-sectional firm-level surveys cover characteristics of 139654 firms from 139 countries. The data is available from https://www.enterprisesurveys.org/. After removing observations with missing values, I analyse 61507 firms operating in 122 countries in this project. The data set of frims used in this project is available at https://github.com/mmchit/edx-project-innovation. To protect the privacy of the firms, I replaced firm ID with a serial number.

To predict an innovative firm, I use seven firm-specific variables plus country, and industry of the firm as predictors. The outcome variable is a dichotomous categorical variable that represents whether the business has introduced a new product or service over the previous three years.

The name of the variables and descriptions are presented in Table 1.

Table 1	
Innovation (outcome)	New products/services introduced over last 3 years
Country	Country where the business operates
Industry	Industry of the business (consolidated to 8 industries)
Age	Years of operation (Survey year – Established year)
Size	Size Category (Small <20; Medium 20-99; Large 100 and above
Legal Status	Firms legal status of incorporation
Foreign Tech	Firm uses technology licensed from a foreign-owned company
Exporter	Firm exports the product
ISO	Firm has an Internationally-Recognized Quality Certification
Training	Firm provides formal training programs for employees

Data and method of analysis

The original data set from the World Bank Enterprise Surveys (WBES) includes 354 variables. To construct the data set used in this analysis, I excluded unnecessary variables and keep only 10 variables used in the analysis. Then, I deleted the observations with missing values. Since the original data set is in STATA format, I did data cleaning in STATA. Then I converted the file to csv format.

The data set used in this project is available for download at https://github.com/mmchit/edx-project-innovation. The following data frame in tidy format shows the structure of the data.

```
# Myint Moe Chit (mmchit@hotmail.com)
# R version: 3.6.0
```

```r
# Download and load required packages (if required)
if(!require(tidyverse)) install.packages("tidyverse",
                                         repos = "http://cran.us.r-project.org")
if(!require(caret)) install.packages("caret",
                                     repos = "http://cran.us.r-project.org")
if(!require(randomForest)) install.packages("randomForest",
                                            repos = "http://cran.us.r-project.org")
if(!require(purrr)) install.packages("purrr",
                                     repos = "http://cran.us.r-project.org")
if(!require(curl)) install.packages("curl",
                                    repos = "http://cran.us.r-project.org")

#Downloading data file from Github

#The full data set is available from the World Bank Enterprise Surveys

#https://www.enterprisesurveys.org/

wbes <- read.csv(curl
                 ("https://raw.githubusercontent.com/mmchit/edx-project-innovation/master/wbes.csv"))

wbes %>% as_tibble()
```

```
## # A tibble: 61,507 x 14
##    country firmId industry size  legalStatus estdYear iso   foreignTech
##    <fct>    <int> <fct>    <fct> <fct>          <int> <fct> <fct>
##  1 Afghan~      1 Manufac~ Medi~ Limited pa~     2006 No    Yes
##  2 Afghan~      2 Manufac~ Smal~ Sole propr~     2012 Yes   No
##  3 Afghan~      3 Manufac~ Smal~ Partnership     2002 No    No
##  4 Afghan~      4 Manufac~ Larg~ Partnership     2007 Yes   No
##  5 Afghan~      5 Manufac~ Medi~ Partnership     1999 No    No
##  6 Afghan~      6 Manufac~ Larg~ Sole propr~     2008 No    No
##  7 Afghan~      7 Manufac~ Medi~ Limited pa~     2008 No    No
##  8 Afghan~      8 Manufac~ Smal~ Partnership     2010 No    Yes
##  9 Afghan~      9 Manufac~ Medi~ Sole propr~     1995 No    No
## 10 Afghan~     10 Manufac~ Medi~ Sole propr~     2007 Yes   No
## # ... with 61,497 more rows, and 6 more variables: innovation <fct>,
## #   employee <int>, training <fct>, surveyYear <int>, age <int>,
## #   exporter <fct>
```

Summary statistics and data visualisation

A brief summary statistics of the key variables and the list of the countries are presented in Table 2.

Table 2

```r
wbes %>% group_by(country) %>%
    summarise(n = n(), Innovation = mean(innovation =="Yes"),
              Avg_Size = mean(employee), Exporter = mean(exporter =="Yes"),
              ISO = mean(iso == "Yes"), Training = mean(training == "Yes"),
              Foreign_Tech = mean(foreignTech == "Yes")) %>%
    knitr::kable(digits = c(0, 0, 2, 2, 2, 2, 2, 2))
```

country	n	Innovation	Avg_Size	Exporter	ISO	Training	Foreign_Tech
Afghanistan	127	0.45	39.15	0.05	0.29	0.40	0.05

country	n	Innovation	Avg_Size	Exporter	ISO	Training	Foreign_Tech
Albania	301	0.11	24.23	0.09	0.25	0.23	0.18
Antiguaandbarbuda	33	0.33	15.12	0.15	0.03	0.30	0.00
Argentina	1969	0.66	141.88	0.20	0.29	0.53	0.14
Armenia	353	0.16	61.35	0.06	0.23	0.19	0.19
Azerbaijan	371	0.02	35.63	0.01	0.13	0.17	0.27
Bahamas	40	0.60	40.90	0.15	0.45	0.48	0.20
Bangladesh	1162	0.36	260.29	0.21	0.23	0.32	0.15
Barbados	63	0.57	42.27	0.37	0.17	0.59	0.11
Belarus	340	0.31	78.84	0.14	0.14	0.48	0.06
Belize	72	0.29	45.53	0.14	0.06	0.36	0.21
Benin	68	0.26	83.18	0.09	0.10	0.21	0.03
Bhutan	75	0.45	35.71	0.19	0.12	0.23	0.13
Bolivia	564	0.72	90.30	0.14	0.20	0.59	0.19
Bosnia and Herzegovina	350	0.37	44.49	0.15	0.31	0.48	0.17
Bulgaria	281	0.24	58.55	0.15	0.24	0.41	0.10
Burundi	57	0.53	73.58	0.12	0.07	0.40	0.12
Cambodia	550	0.32	261.44	0.15	0.11	0.60	0.18
Cameroon	88	0.38	73.19	0.15	0.09	0.33	0.20
Centralafricanrepublic	29	0.38	49.90	0.07	0.34	0.41	0.24
Chad	72	0.33	21.29	0.01	0.01	0.33	0.08
Chile	1297	0.63	120.70	0.16	0.31	0.54	0.16
China	1615	0.47	291.79	0.17	0.72	0.86	0.24
Colombia	1769	0.66	107.65	0.14	0.23	0.58	0.10
Costarica	288	0.63	93.45	0.21	0.18	0.53	0.08
Côte d'Ivoire	97	0.34	88.55	0.20	0.14	0.29	0.05
Croatia	339	0.39	54.97	0.18	0.28	0.51	0.16
Czech Republic	234	0.50	118.11	0.29	0.39	0.53	0.15
Djibouti	242	0.34	35.01	0.14	0.17	0.21	0.19
Dominica	28	0.14	29.86	0.32	0.11	0.21	0.11
DominicanRepublic	210	0.47	175.87	0.23	0.17	0.52	0.26
DRC	227	0.44	42.88	0.04	0.15	0.15	0.07
Ecuador	531	0.73	104.67	0.11	0.24	0.67	0.17
Egypt	3830	0.17	149.41	0.13	0.25	0.15	0.09
ElSalvador	902	0.51	123.36	0.26	0.15	0.47	0.15
Estonia	240	0.22	30.02	0.21	0.18	0.38	0.12
Eswatini	59	0.22	95.22	0.29	0.19	0.19	0.07
Ethiopia	562	0.44	155.20	0.09	0.15	0.26	0.23
Gambia	58	0.47	18.03	0.10	0.19	0.24	0.14
Georgia	357	0.10	34.57	0.04	0.12	0.11	0.16
Ghana	351	0.53	36.62	0.08	0.10	0.35	0.15
Greece	563	0.26	60.96	0.25	0.64	0.30	0.18
Grenada	17	0.76	30.41	0.12	0.29	0.41	0.12
Guatemala	763	0.62	127.27	0.23	0.14	0.48	0.18
Guinea	23	0.22	62.13	0.09	0.17	0.17	0.22
Guyana	66	0.39	144.26	0.27	0.30	0.52	0.17
Honduras	451	0.55	92.38	0.14	0.19	0.43	0.17
Hungary	289	0.21	63.85	0.11	0.54	0.16	0.07
India	7004	0.45	106.51	0.12	0.49	0.41	0.10
Indonesia	1040	0.13	160.58	0.14	0.22	0.11	0.27
Israel	474	0.24	111.69	0.16	0.37	0.17	0.06
Jamaica	109	0.37	92.38	0.17	0.21	0.35	0.17
Jordan	514	0.22	109.15	0.29	0.16	0.05	0.10

country	n	Innovation	Avg_Size	Exporter	ISO	Training	Foreign_Tech
Kazakhstan	572	0.19	65.15	0.02	0.19	0.33	0.09
Kenya	798	0.56	121.86	0.21	0.32	0.43	0.20
Kosovo	191	0.52	29.22	0.08	0.33	0.53	0.34
Kyrgyz Republic	258	0.39	56.16	0.09	0.25	0.55	0.10
LaoPDR	245	0.20	63.18	0.17	0.06	0.19	0.14
Latvia	295	0.21	31.14	0.20	0.20	0.30	0.10
Lebanon	540	0.42	53.92	0.26	0.21	0.22	0.07
Lesotho	67	0.06	315.46	0.21	0.19	0.43	0.31
Liberia	74	0.57	33.70	0.08	0.00	0.30	0.07
Lithuania	241	0.25	42.90	0.20	0.21	0.37	0.18
Malawi	149	0.55	108.01	0.07	0.20	0.16	0.25
Malaysia	505	0.13	222.94	0.42	0.38	0.32	0.22
Mali	78	0.33	42.10	0.21	0.04	0.18	0.12
Mauritania	46	0.43	74.02	0.39	0.17	0.63	0.09
Mexico	2073	0.42	142.43	0.12	0.22	0.44	0.13
Moldova	333	0.30	42.87	0.07	0.17	0.35	0.17
Mongolia	339	0.27	49.26	0.04	0.14	0.58	0.15
Montenegro	140	0.11	37.32	0.09	0.19	0.20	0.11
Morocco	342	0.29	93.72	0.17	0.25	0.30	0.16
Mozambique	280	0.34	54.46	0.13	0.10	0.16	0.21
Myanmar	698	0.24	123.98	0.13	0.04	0.11	0.05
Namibia	144	0.55	43.76	0.12	0.15	0.35	0.18
Nepal	238	0.41	72.80	0.11	0.16	0.27	0.05
Nicaragua	565	0.50	59.23	0.09	0.16	0.37	0.10
Niger	40	0.40	23.15	0.00	0.10	0.18	0.05
Nigeria	908	0.55	38.07	0.17	0.10	0.28	0.12
North Macedonia	349	0.31	34.60	0.12	0.32	0.44	0.15
Pakistan	845	0.32	185.91	0.14	0.35	0.22	0.18
Panama	321	0.41	50.44	0.07	0.13	0.32	0.13
PapuaNewGuinea	21	0.43	162.71	0.00	0.24	0.71	0.33
Paraguay	583	0.67	73.63	0.15	0.14	0.52	0.13
Peru	1610	0.70	163.31	0.23	0.25	0.67	0.12
Philippines	829	0.36	99.10	0.22	0.25	0.53	0.16
Poland	505	0.33	62.88	0.14	0.32	0.34	0.15
Romania	521	0.41	47.73	0.14	0.35	0.44	0.14
Russia	3995	0.25	64.60	0.03	0.11	0.43	0.08
Rwanda	60	0.58	81.07	0.08	0.22	0.57	0.22
Senegal	226	0.44	37.28	0.09	0.07	0.11	0.13
Serbia	335	0.35	84.24	0.19	0.37	0.34	0.16
SierraLeone	76	0.34	23.12	0.04	0.05	0.26	0.09
Slovak Republic	228	0.20	38.33	0.18	0.44	0.43	0.30
Slovenia	254	0.34	103.00	0.29	0.27	0.47	0.16
Solomon Islands	31	0.35	92.10	0.39	0.10	0.45	0.26
Southsudan	89	0.70	24.12	0.02	0.07	0.16	0.26
SriLanka	347	0.24	102.12	0.07	0.14	0.22	0.09
StKittsandNevis	28	0.43	54.39	0.36	0.11	0.39	0.14
StLucia	63	0.16	25.24	0.29	0.03	0.33	0.00
StVincentandGrenadines	45	0.47	30.82	0.27	0.22	0.33	0.29
Sudan	61	0.61	38.44	0.02	0.10	0.13	0.10
Suriname	132	0.55	27.65	0.17	0.21	0.17	0.05
Sweden	290	0.77	149.19	0.53	0.72	0.66	0.20
Tajikistan	327	0.16	36.73	0.05	0.19	0.30	0.17

country	n	Innovation	Avg_Size	Exporter	ISO	Training	Foreign_Tech
Tanzania	384	0.57	68.34	0.08	0.21	0.29	0.17
Thailand	581	0.09	125.92	0.22	0.31	0.39	0.10
Timor-Leste	57	0.46	17.75	0.40	0.02	0.00	0.23
Togo	44	0.45	109.93	0.59	0.25	0.25	0.07
TrinidadandTobago	123	0.46	74.54	0.24	0.24	0.35	0.07
Tunisia	549	0.26	95.80	0.34	0.24	0.32	0.07
Turkey	1220	0.12	112.09	0.30	0.46	0.37	0.31
Uganda	315	0.71	63.89	0.10	0.21	0.31	0.22
Ukraine	895	0.20	53.87	0.09	0.18	0.20	0.13
Uruguay	759	0.65	58.74	0.22	0.15	0.41	0.10
Uzbekistan	383	0.05	127.88	0.04	0.12	0.25	0.10
Venezuela	69	0.36	85.13	0.04	0.14	0.48	0.07
Vietnam	612	0.35	230.46	0.24	0.24	0.25	0.12
West Bank And Gaza	409	0.20	22.93	0.15	0.12	0.14	0.09
Yemen	336	0.40	42.45	0.05	0.14	0.25	0.12
Zambia	339	0.54	38.18	0.06	0.20	0.26	0.21
Zimbabwe	588	0.48	99.39	0.09	0.27	0.29	0.16

The number of innovative business also varies across different industries. Following Table 3 shows the distribution of innovative firms and selected firm-specific characteristics across different industries.

Table 3

```r
wbes %>% group_by(industry) %>%
    summarise(n = n(), Innovation = mean(innovation =="Yes"),
              Avg_Size = mean(employee), Exporter = mean(exporter =="Yes"),
              ISO = mean(iso == "Yes"), Training = mean(training == "Yes"),
              Foreign_Tech = mean(foreignTech == "Yes")) %>%
    knitr::kable(digits = c(0, 0, 2, 2, 2, 2, 2, 2))
```

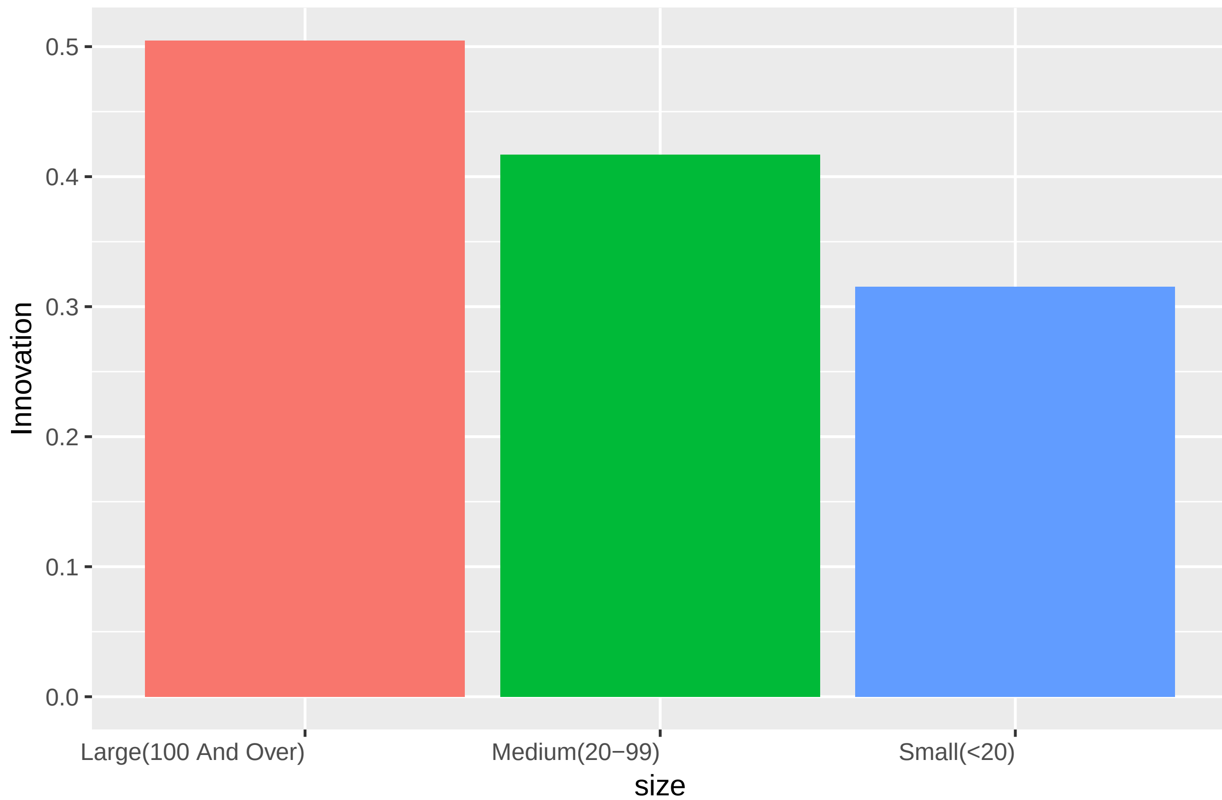
industry	n	Innovation	Avg_Size	Exporter	ISO	Training	Foreign_Tech
Construction	559	0.19	81.92	0.00	0.16	0.48	0.07
Food	7418	0.45	142.88	0.15	0.28	0.42	0.13
Garments & Textile	6600	0.44	175.31	0.23	0.19	0.34	0.14
Manufacturing	33473	0.43	112.43	0.17	0.32	0.39	0.15
Others	1473	0.67	55.15	0.15	0.14	0.43	0.13
Retail & Wholesale	5790	0.18	46.49	0.04	0.13	0.32	0.08
Services	5860	0.24	49.22	0.09	0.21	0.37	0.14
Transport	334	0.10	69.63	0.11	0.17	0.31	0.07

Note Except the average size of firms in the number of employees, the remaining variables are measured in proportion.

To visually evaluate the relationship between the predictors and the outcome variable, I created a number of plots.

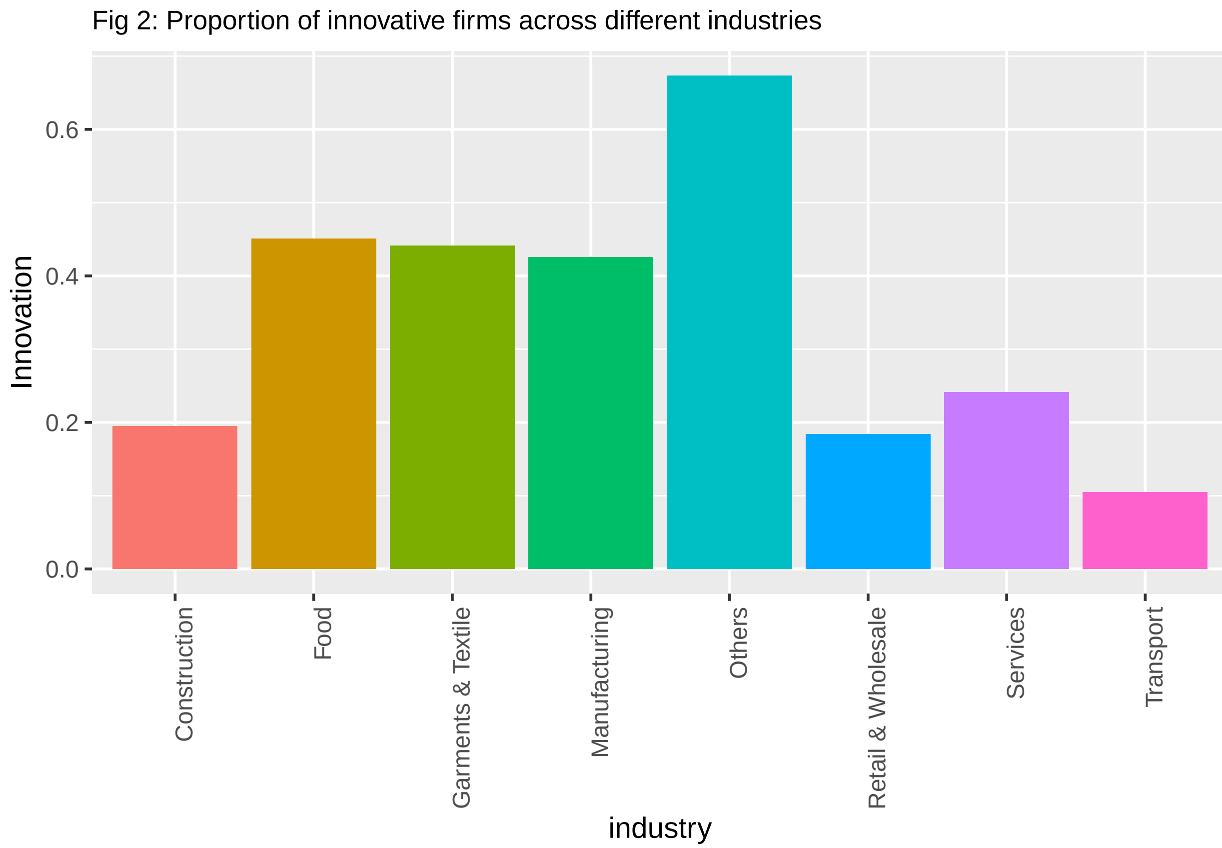
```r
wbes %>% group_by(size) %>% summarize(Innovation = mean(innovation == "Yes")) %>%
    ggplot(aes(size, Innovation)) + geom_bar(stat = "identity", aes(fill = size)) +
    theme(axis.text.x = element_text(hjust = 1), legend.position = "none",
          plot.title = element_text(size = 10, face = "plain")) +
  ggtitle ("Fig 1: Proportion of innovative firms across different sizes")
```

Fig 1: Proportion of innovative firms across different sizes



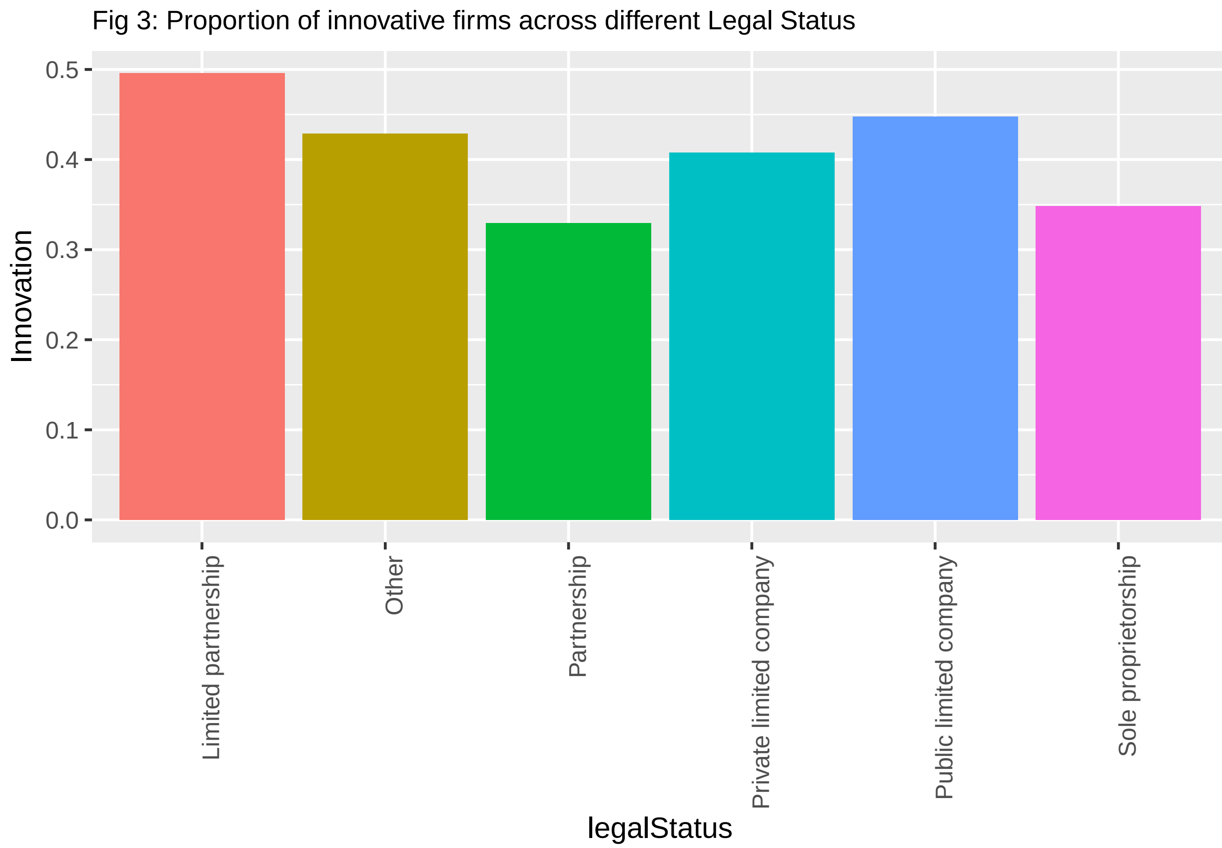
As shown in Figure 1, the proportion of innovative frims are different across size-categories.

```
wbes %>% group_by(industry) %>% summarize(Innovation = mean(innovation == "Yes")) %>%
    ggplot(aes(industry, Innovation)) + geom_bar(stat = "identity", aes(fill = industry)) +
    theme(axis.text.x = element_text(angle = 90, hjust = 1),
          plot.title = element_text(size = 10, face = "plain"), legend.position = "none") +
  ggtitle ("Fig 2: Proportion of innovative firms across different industries")
```



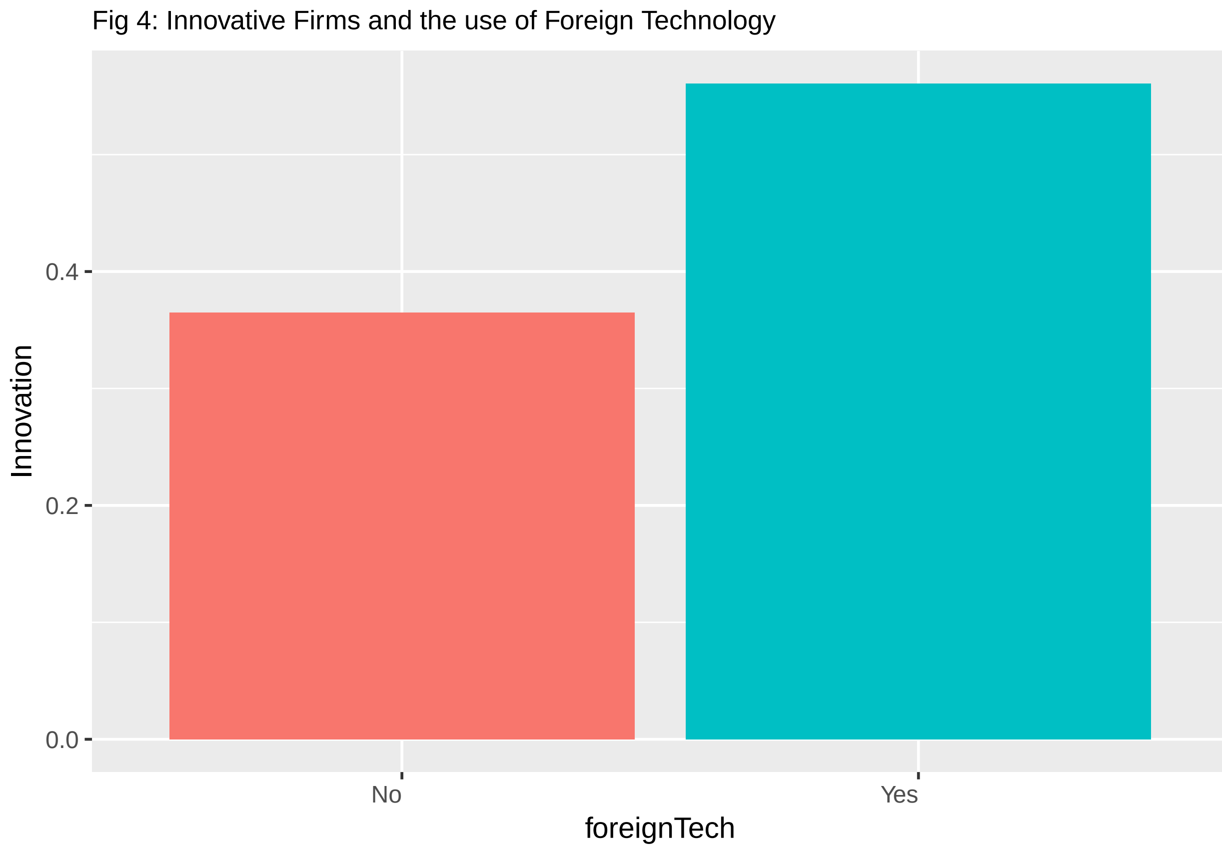
Firms' capability to innovate will also different across the industry in which they operate. Manufacturing firms are more likely to offer new products compared to firms in construction industry. I plot the proportion of innovative firms across industries and presented in Figure 2.

```
wbes %>% group_by(legalStatus) %>% summarize(Innovation = mean(innovation == "Yes")) %>%
    ggplot(aes(legalStatus, Innovation)) + geom_bar(stat = "identity", aes(fill = legalStatus)) +
    theme(axis.text.x = element_text(angle = 90, hjust = 1),
          plot.title = element_text(size = 10, face = "plain"), legend.position = "none") +
  ggtitle ("Fig 3: Proportion of innovative firms across different Legal Status")
```



A firm's legal status could also be associated with its innovation. Figure 3 shows the difference in the proportion of innovative firms across different types of legal status.

```
wbes %>% group_by(foreignTech) %>% summarize(Innovation = mean(innovation == "Yes")) %>%
    ggplot(aes(foreignTech, Innovation)) + geom_bar(stat = "identity", aes(fill = foreignTech)) +
    theme(axis.text.x = element_text(hjust = 1),
          plot.title = element_text(size = 10, face = "plain"),
          legend.position = "none") +
  ggtitle ("Fig 4: Innovative Firms and the use of Foreign Technology")
```



It is reasonable to assume that firms that use foreign technology are more likely to offer new and innovative products.

As shown in the figure, more than 55% of firms that use licensed foreign technology introduce a new product or servide. In comparison, only 36% of firms that do not have access to foreign technology introduce a new product.

```
wbes %>% group_by(exporter) %>% summarize(Innovation = mean(innovation == "Yes")) %>%
    ggplot(aes(exporter, Innovation)) + geom_bar(stat = "identity", aes(fill = exporter)) +
    theme(axis.text.x = element_text(hjust = 1),
          plot.title = element_text(size = 10, face = "plain"),
          legend.position = "none") +
  ggtitle ("Fig 5: Innovative Firms and the Exporters")
```

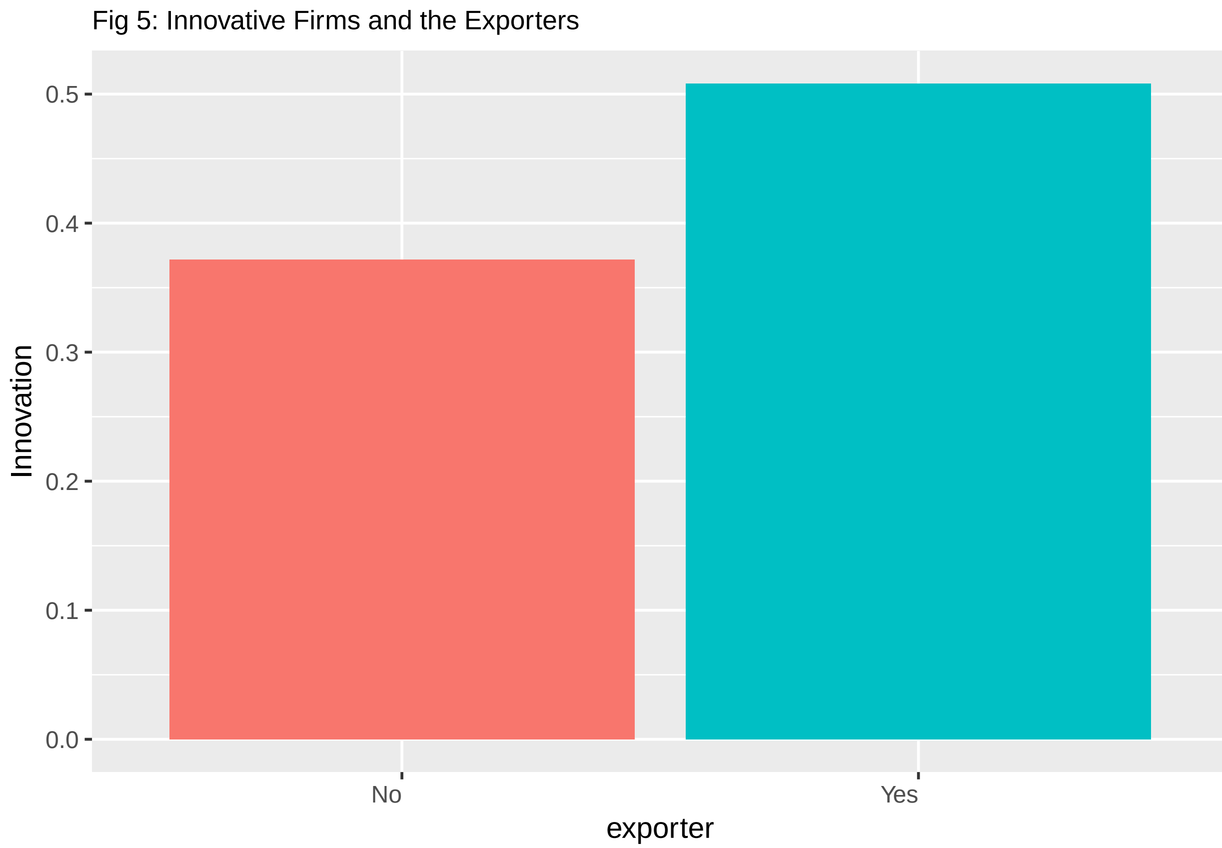


Figure 5 shows the difference in the proportion of innovative frims between expoters are non-exporters.

```
wbes %>% group_by(iso) %>% summarize(Innovation = mean(innovation == "Yes")) %>%
    ggplot(aes(iso, Innovation)) + geom_bar(stat = "identity", aes(fill = iso)) +
    theme(axis.text.x = element_text(hjust = 1),
          plot.title = element_text(size = 10, face = "plain"),
          legend.position = "none") +
  ggtitle ("Fig 6: Innovative Firms and ISO certificate")
```

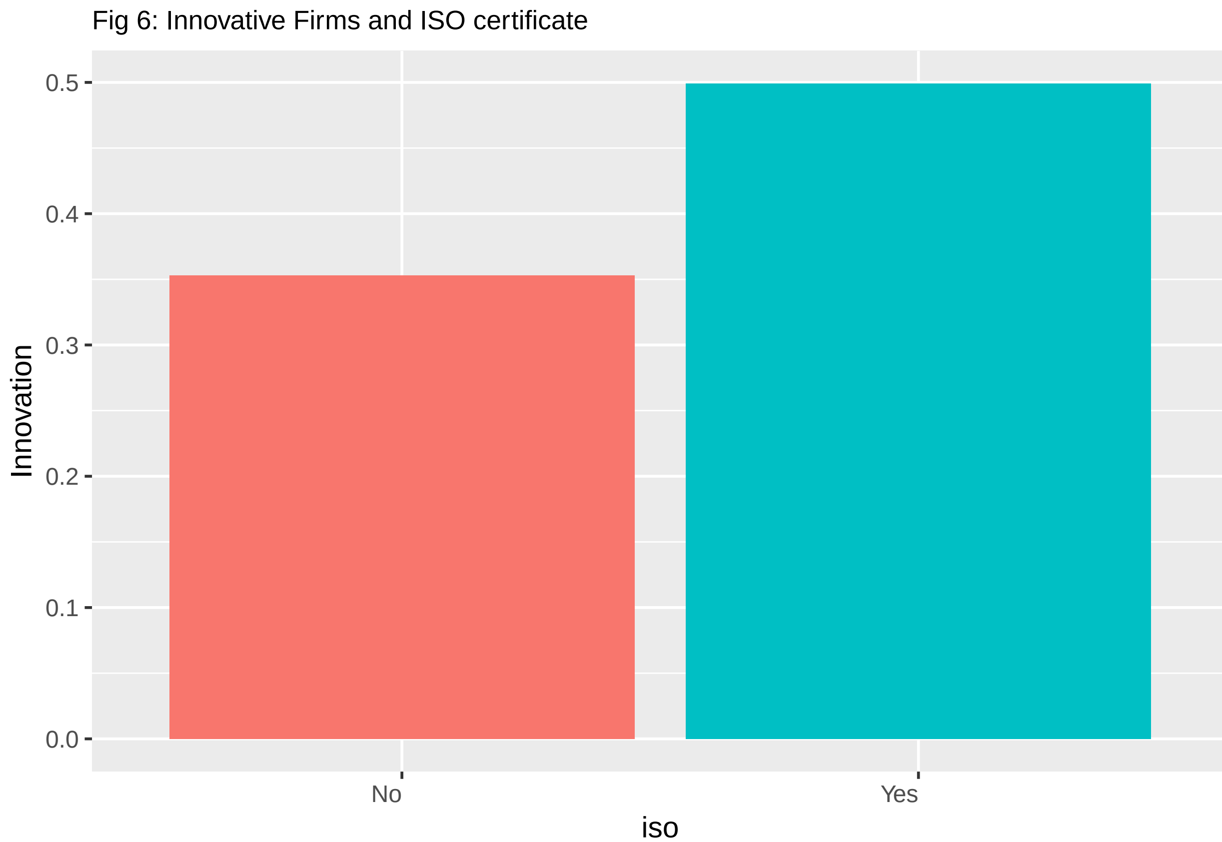


Figure 6 demonstrates that firms with internationally recognized certification, such as ISO9001, are more likely to be innovative.

```
wbes %>% group_by(training) %>% summarize(Innovation = mean(innovation == "Yes")) %>%
    ggplot(aes(training, Innovation)) + geom_bar(stat = "identity", aes(fill = training)) +
    theme(axis.text.x = element_text(hjust = 1),
          plot.title = element_text(size = 10, face = "plain"),
          legend.position = "none") +
  ggtitle ("Fig 7: Innovative Firms and Training")
```

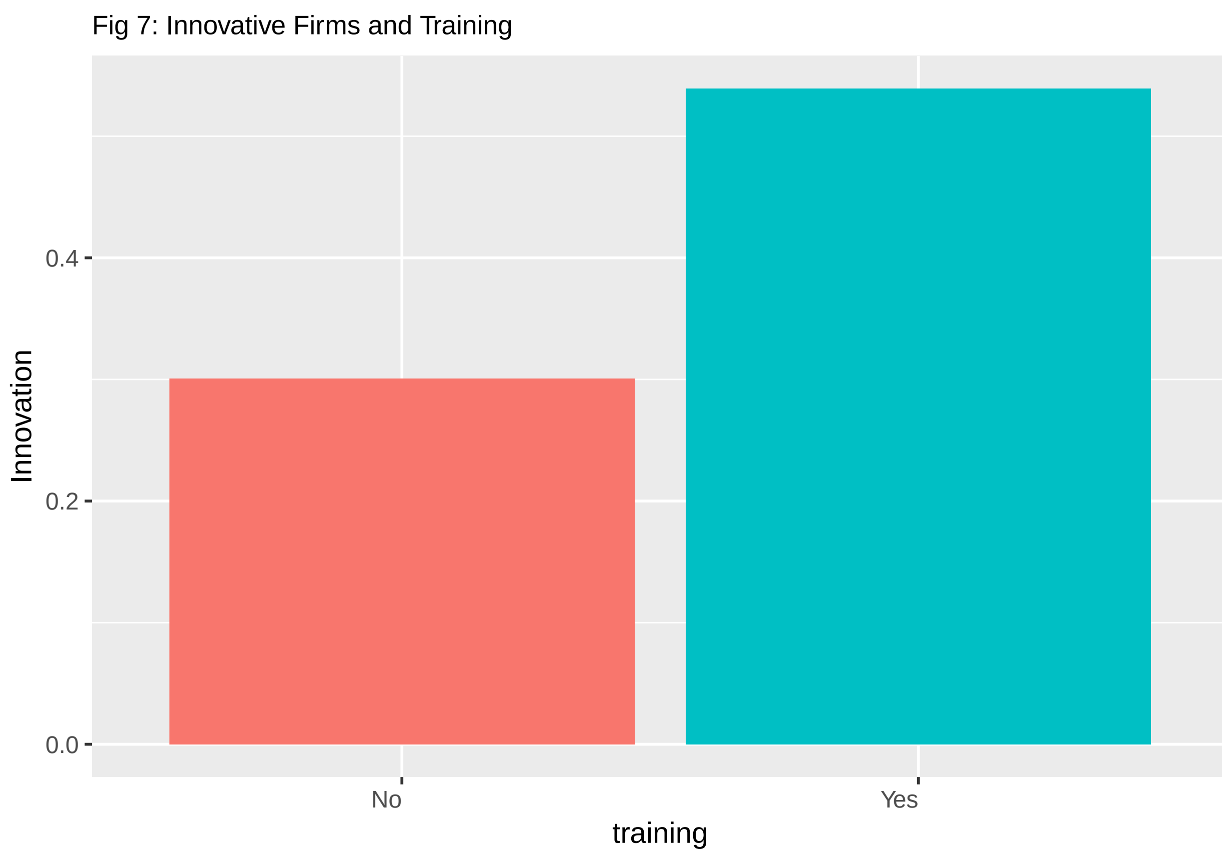


Figure 7 also shows that firms that provide training to their employees are more innovative.

Machine learning algorithms

As presented tables and figures suggest, the predictors used in this project are significantly associated with a firm's capabilities to innovate and offer a new product or service.

It is important to select a machine learning algorithm suitable for the type and distribution of data. Since the outcome is a binary categorical variable and the model includes nine predictors, linear regression and Quadratic Discriminant Analysis are considered not suitable. The random forest algorithm is also not suitable because one of the predictors, country, consists of 122 factors. Therefore, I use the following machine learning algorithms.

- Simple random selection (benchmark)
- Logitic regression
- k-nearest neighbours
- Naïve Bayes
- Linear discriminant analysis
- Classification tree
- Random forest (Rborist)

I will select the algorithm with the highest accuracy and the lowest residual mean squared error (although RMSE may not very informative for a categorical outcome variable).

Training and test data set

I divide the data set into a training set (90% of the observations) with 55355 observations and a test set (10% of the observations) with 6125 observations.

```r
#Creating a training set (90%) and a test set (10%)

y <- wbes$innovation

mean(y == "Yes")
```

```
## [1] 0.3920529
```

```r
set.seed(1)
test_index <- createDataPartition(y, times = 1, p = 0.1, list = FALSE)
test_set <- wbes[test_index, ]
train_set <- wbes[-test_index, ]
```

Analysis and results

The cretirea for selecting the most appropriate machine learning algorithm are accuracy and the residual mean squared error. The residual mean squared error is calculated using the following function.

```r
RMSE <- function(true_ratings, predicted_ratings){
    sqrt(mean((true_ratings - predicted_ratings)^2))
}
```

Baseline model (Random selection) As a banch mark, I first create a simple model based on random selection.It is not surprising, on average, about 50 percent of firms are correctly predicted using random selection.

```r
#Baseline prediction: Selecting firms in random

y_hat <- sample(c("Yes", "No"), length(test_index), replace = TRUE) %>%
    factor(levels = levels(test_set$innovation))

mean(y_hat == test_set$innovation)
```

```
## [1] 0.4936606
```

```r
confusionMatrix(data = y_hat, reference = test_set$innovation, positive = "Yes")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  1820 1195
##        Yes 1920 1217
##
##                Accuracy : 0.4937
##                  95% CI : (0.4811, 0.5062)
##     No Information Rate : 0.6079
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : -0.0084
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.5046
##             Specificity : 0.4866
##          Pos Pred Value : 0.3880
```

```
##          Neg Pred Value : 0.6036
##              Prevalence : 0.3921
##          Detection Rate : 0.1978
##    Detection Prevalence : 0.5099
##       Balanced Accuracy : 0.4956
##
##        'Positive' Class : Yes
##
```

```
accuracy_random <- confusionMatrix(data = y_hat, reference = test_set$innovation,
                                   positive = "Yes")$overall["Accuracy"]

rmse_random <- RMSE(as.numeric(test_set$innovation), as.numeric(y_hat))

Results <- data_frame(Method = "Random Selection", Accuracy = accuracy_random,
                      RMSE = rmse_random)
```

As shown in the table, The accuracy is 0.49 and RMSE is 0.71 (closer to 1).

```
Results %>% knitr::kable(digits = c(0, 2, 2))
```

Method	Accuracy	RMSE
Random Selection	0.49	0.71

Logitic Regression As a first machine learning algrothm, I use logistic regression, wich is a generalised linear model suitable for a binary outcome variable.

```
#Logit Regression

logit_fit <- glm(innovation ~ age + country + size + legalStatus + industry + foreignTech
                 + exporter + iso + training, data = train_set, family = "binomial")

p_h_logit <- predict(logit_fit, test_set)

y_h_logit <- factor(ifelse(p_h_logit > 0.5, "Yes", "No"))

confusionMatrix(data = y_h_logit, reference = test_set$innovation, positive = "Yes")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  3474 1635
##        Yes  266  777
##
##                Accuracy : 0.691
##                  95% CI : (0.6793, 0.7025)
##     No Information Rate : 0.6079
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.2791
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.3221
```

```
##             Specificity : 0.9289
##          Pos Pred Value : 0.7450
##          Neg Pred Value : 0.6800
##              Prevalence : 0.3921
##          Detection Rate : 0.1263
##    Detection Prevalence : 0.1695
##       Balanced Accuracy : 0.6255
##
##        'Positive' Class : Yes
##
```

```
accuracy_logit <- confusionMatrix(data = y_h_logit, reference = test_set$innovation,
                                  positive = "Yes")$overall["Accuracy"]

rmse_logit <- RMSE(as.numeric(test_set$innovation), as.numeric(y_h_logit))

Results1 <- data.frame(Method = "Logitic Regression",
                       Accuracy = accuracy_logit, RMSE = rmse_logit)
```

The accuracy and RMSE obtained from Logistic regression are significantly higher than those from simple random selection.

```
Results1 %>% knitr::kable(digits = c(0, 2, 2))
```

	Method	Accuracy	RMSE
Accuracy	Logitic Regression	0.69	0.56

K-nearest neighbours I then use K-nearest neighbours algorithm that pay attention to the relationship and trend among the observations close to each other. This method allow us to control the flexibility of the estimates by setting the number of the points in the neighbourhood used to compute the average (Irzarry 2019). After employing a cross-validation method (not presented here), the optimum number of k that maximise the accuracy of the model is 13. The accuracy of this method is lower than LOgistic regression.

```
#k-Nearest Neigbours

knn_fit <- knn3(innovation ~ age + country + size + legalStatus + industry + foreignTech
                + exporter + iso + training, data = train_set, k = 13)

y_h_knn <- predict(knn_fit, test_set, type = "class")

confusionMatrix(data = y_h_knn, reference = test_set$innovation, positive = "Yes")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  3097 1392
##        Yes  643 1020
##
##                Accuracy : 0.6692
##                  95% CI : (0.6573, 0.681)
##     No Information Rate : 0.6079
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.2656
```

```
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4229
##             Specificity : 0.8281
##          Pos Pred Value : 0.6133
##          Neg Pred Value : 0.6899
##              Prevalence : 0.3921
##          Detection Rate : 0.1658
##    Detection Prevalence : 0.2703
##       Balanced Accuracy : 0.6255
##
##        'Positive' Class : Yes
##
```

```
accuracy_knn <- confusionMatrix(data = y_h_knn, reference = test_set$innovation,
                positive = "Yes")$overall["Accuracy"]

rmse_knn <- RMSE(as.numeric(test_set$innovation), as.numeric(y_h_knn))

Results1 <- data.frame(Method = "k-Nearest Neigbour",
                                 Accuracy = accuracy_knn, RMSE = rmse_knn)
```

```
Results1 %>% knitr::kable(digits = c(0, 2, 2))
```

	Method	Accuracy	RMSE
Accuracy	k-Nearest Neigbour	0.67	0.58

**Naive Bayes method Although this algorithm is not suitable for a model with more than two predictors, I estimate the model with Naive Bayes method, which is one of the generative models, that based on the joint distribution of outcome and predictors. As we suspect, the accuracy of the model obtained from using Naive Bayes method is much lower.

```
# Naive Bayes

naive_fit <- train(innovation ~ age + country + size + legalStatus + industry + foreignTech
                   + exporter + iso + training, data = train_set, method = "naive_bayes")

y_h_naive <- predict(naive_fit, test_set)

confusionMatrix(data = y_h_naive, reference = test_set$innovation, positive = "Yes")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  3739 2410
##        Yes    1    2
##
##                Accuracy : 0.6081
##                  95% CI : (0.5958, 0.6203)
##     No Information Rate : 0.6079
##     P-Value [Acc > NIR] : 0.4952
##
```

```
##                  Kappa : 7e-04
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.0008292
##             Specificity : 0.9997326
##          Pos Pred Value : 0.6666667
##          Neg Pred Value : 0.6080664
##              Prevalence : 0.3920676
##          Detection Rate : 0.0003251
##    Detection Prevalence : 0.0004876
##       Balanced Accuracy : 0.5002809
##
##        'Positive' Class : Yes
##
```

```r
accuracy_naive <- confusionMatrix(data = y_h_naive, reference = test_set$innovation,
                                  positive = "Yes")$overall["Accuracy"]

rmse_naive <- RMSE(as.numeric(test_set$innovation), as.numeric(y_h_naive))

Results1 <- data.frame(Method = "Naive Bayes",
                                  Accuracy = accuracy_naive, RMSE = rmse_naive)
```

```r
Results1 %>% knitr::kable(digits = c(0, 2, 2))
```

	Method	Accuracy	RMSE
Accuracy	Naive Bayes	0.61	0.63

Linear Discriminant Analysis method

One of the generative models that provides a relatively simple solution to the problem of having too many parameters is Linear Discriminant Analysis (LDA). This method assumes that the correlation structure of the predictors is the same for all classes, which reduces the number of parameters we need to estimate (Irzarry 2019). Since the model used in this project includes nine predictors, LDA is considered to be a thoeretically appropriate method. The obtained accuracy and RMSE confirm this. So far, this method produces the higher accuracy and the lowest RMSE.

```r
# Linear Discriminant Analysis

lda_fit <- train(innovation ~ age + country + size + legalStatus +
                     industry +  foreignTech + exporter + iso +
                     training , data = train_set, method = "lda")

y_h_lda <- predict(lda_fit, test_set)

confusionMatrix(data = y_h_lda, reference = test_set$innovation, positive = "Yes")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  3075 1124
##        Yes  665 1288
##
```

```
##                Accuracy : 0.7092
##                  95% CI : (0.6977, 0.7205)
##     No Information Rate : 0.6079
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3686
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.5340
##             Specificity : 0.8222
##          Pos Pred Value : 0.6595
##          Neg Pred Value : 0.7323
##              Prevalence : 0.3921
##          Detection Rate : 0.2094
##    Detection Prevalence : 0.3175
##       Balanced Accuracy : 0.6781
##
##        'Positive' Class : Yes
##
```

```r
accuracy_lda <- confusionMatrix(data = y_h_lda, reference = test_set$innovation,
                                positive = "Yes")$overall["Accuracy"]

rmse_lda <- RMSE(as.numeric(test_set$innovation), as.numeric(y_h_lda))

Results1 <- data.frame(Method = "Linear Descriminant Analysis",
                                Accuracy = accuracy_lda, RMSE = rmse_lda)
```

```r
Results1 %>% knitr::kable(digits = c(0, 2, 2))
```

	Method	Accuracy	RMSE
Accuracy	Linear Descriminant Analysis	0.71	0.54

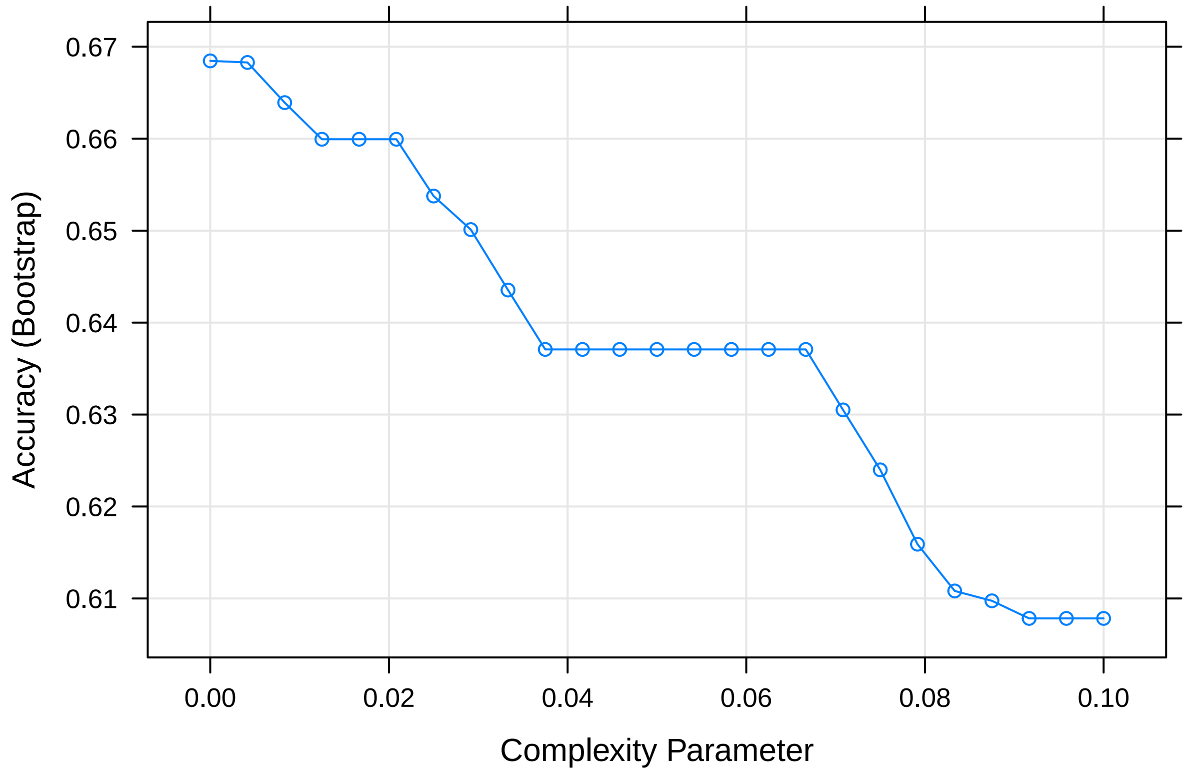
Classification (decision) Trees

Another way to overcome the problem associated with using many predictors is to use methods that allow higher dimensions in predictor variables. Classification trees, or decision trees, method is one of the methods used in prediction problems where the categorical outcome is associated with many predictors (Irzarry 2019). However, this method scores accuracy lower than LDA method.

```r
#Classification Trees

train_rpart <- train(innovation ~ age + country + size + legalStatus + iso
                     + industry +  foreignTech + exporter + training,
                     method = "rpart",
                     tuneGrid = data.frame(cp = seq(0.0, 0.1, len = 25)),
                     data = train_set)

plot(train_rpart)
```




```r
y_h_rpart <- predict(train_rpart, test_set)

confusionMatrix(data = y_h_rpart, reference = test_set$innovation, positive = "Yes")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  3044 1204
##        Yes  696 1208
##
##                Accuracy : 0.6912
##                  95% CI : (0.6794, 0.7027)
##     No Information Rate : 0.6079
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.327
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.5008
##             Specificity : 0.8139
##          Pos Pred Value : 0.6345
##          Neg Pred Value : 0.7166
##              Prevalence : 0.3921
##          Detection Rate : 0.1964
```

```
##    Detection Prevalence : 0.3095
##       Balanced Accuracy : 0.6574
##
##        'Positive' Class : Yes
##
```

```r
accuracy_rpart <- confusionMatrix(data = y_h_rpart, reference = test_set$innovation,
                                  positive = "Yes")$overall["Accuracy"]

rmse_rpart <- RMSE(as.numeric(test_set$innovation), as.numeric(y_h_rpart))

Results1 <- data.frame(Method = "Classification Trees",
                                         Accuracy = accuracy_rpart, RMSE = rmse_rpart)
```

```r
Results1 %>% knitr::kable(digits = c(0, 2, 2))
```

	Method	Accuracy	RMSE
Accuracy	Classification Trees	0.69	0.56

Random forest (Rborist) method

Random forests approach uses a method to improve prediction performance and reduce instability by averaging multiple decision trees. This method is able to generate many predictors, each using regression or classification trees, and then forming a final prediction based on the average prediction of all these trees (Irzarry 2019). As shown in the table, the accuracy obtained from this model is also lower than the one from LDA.

```r
#Random Forest Rborist
# This analysis takes about an hour

train_rb <- train(innovation ~ age + country + size + legalStatus + iso
                 + industry +  foreignTech + exporter +
                     training , data = train_set, method = "Rborist")

y_h_rb <- predict(train_rb, test_set)

confusionMatrix(data = y_h_rb, reference = test_set$innovation, positive = "Yes")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  3558 2184
##        Yes  182  228
##
##                Accuracy : 0.6154
##                  95% CI : (0.6031, 0.6276)
##     No Information Rate : 0.6079
##     P-Value [Acc > NIR] : 0.1173
##
##                   Kappa : 0.0538
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.09453
##             Specificity : 0.95134
```

```
##          Pos Pred Value : 0.55610
##          Neg Pred Value : 0.61964
##              Prevalence : 0.39207
##          Detection Rate : 0.03706
##    Detection Prevalence : 0.06664
##       Balanced Accuracy : 0.52293
##
##        'Positive' Class : Yes
##
```

```
accuracy_rb <- confusionMatrix(data = y_h_rb, reference = test_set$innovation,
                                 positive = "Yes")$overall["Accuracy"]

rmse_rb <- RMSE(as.numeric(test_set$innovation), as.numeric(y_h_rb))

Results1 <- data.frame(Method = "Random Forest",
                        Accuracy = accuracy_rb, RMSE = rmse_rb)
```

```
Results1 %>% knitr::kable(digits = c(0, 2, 2))
```

	Method	Accuracy	RMSE
Accuracy	Random Forest	0.62	0.62

Summary of the findings

Based on the values of accuracy and RMSE obtained from the models, Linear Disccriminat Analysis produces the highest accuracy score and the lowest RMSE. The results indicate that the machine learning model developed in this project can predict whether a frim is innovative or not at about 71 percent of accuracy.

```
#Summary of results

Results <- bind_rows(Results, data.frame(Method = "Logit Regression",
                                Accuracy = accuracy_logit, RMSE = rmse_logit))

Results <- bind_rows(Results, data.frame(Method = "k-Nearest Neigbours",
                                Accuracy = accuracy_knn, RMSE = rmse_knn))

Results <- bind_rows(Results, data.frame(Method = "Naive Bayes",
                                         Accuracy = accuracy_naive, RMSE = rmse_naive))

Results <- bind_rows(Results, data.frame(Method = "Linear Discriminant Analysis",
                                         Accuracy = accuracy_lda, RMSE = rmse_lda))

Results <- bind_rows(Results, data.frame(Method = "Classification Trees",
                                         Accuracy = accuracy_rpart, RMSE = rmse_rpart))

Results <- bind_rows(Results, data.frame(Method = "Random Forest",
                                         Accuracy = accuracy_rb, RMSE = rmse_rb))

Results %>% knitr::kable(digits = c(0, 2, 2))
```

Method	Accuracy	RMSE
Random Selection	0.49	0.71
Logit Regression	0.69	0.56

Method	Accuracy	RMSE
k-Nearest Neigbours	0.67	0.58
Naive Bayes	0.61	0.63
Linear Discriminant Analysis	0.71	0.54
Classification Trees	0.69	0.56
Random Forest	0.62	0.62

Variable Importance

To identify the relative importance of predictors in estimating whether a frim is innovative or not, I obtain the variable importance scores using Linear Discriminant Analysis. The scores are as follows:

```
#Variable Importance

varImp(lda_fit, scale = FALSE)
```

```
## ROC curve variable importance
##
##              Importance
## training        0.6179
## size            0.5834
## industry        0.5668
## age             0.5656
## iso             0.5603
## foreignTech     0.5486
## country         0.5436
## exporter        0.5359
## legalStatus     0.5305
```

The scores clearly indiccate that providing training (an indicator of capacity building), the size of firm (an indicator for managerial and financial resources), and frim age (an indicator for experience) are relatively more important for firms' innovative capability.

Conclusion

In this project, I attempted to predict an innovative firm using a data set obtained from the World Bank Enterprises Surveys. Based on summary statistics and data visualisation plots, I provided justification for the predictors used in the model. The accuracy of the estimates obtained from Linear Discriminant Analysis method indicate that the machine learning model can predict just over 70 percent of firms correctly.

The importance of variable obtained from Linear Discriminant Analysis is also in accordance with business theories. The most importance predictors for an innovative firm are training,

Given the model is based on only nine predictors, the obtained accuracy is reasonable. The model can be further improved by adding more firm-specific variables, such as the experience of the manager, the level of education of employees, and the level of institutional development of the country.

Reference:

Irzarry (2019) Introduction to Data Science: Data Analysis and Prediction Algorithms with R (available at: https://rafalab.github.io/dsbook/)

Note The estimation is conducted using R3.6.0.