Advanced Business Analytics – Power of Predictive Modelling

Wednesday 17:10 Course code: 226161-1380

Course Instructor: dr Adam Celestyn Korczyński

Practical Application of Multiple Imputation Estimation for Predicting Sales

Winter Term 2021/2022

Group:

Steve Mocanu 110054 – Introduction, dataset description and quality, theoretical overview

Jose Caloca Martinez 110558 **(Group Leader)** – Variable selection (Random Forest), Model after imputation, conclusion

Diego Rodriguez Martinez 110074 – Model before imputation

Wiktor Szczepanowski 108772 – descriptive statistics, exploratory data analysis

Prasenjeet Rathore 110092 – Multiple imputation, conclusion

1. Introduction

Our project focuses on gaining valuable insights from the joint EBRD-EIB-WB Enterprise Surveys conducted in 2018-2020. This collection of data from enterprises in manufacturing and service sectors was made possible by the collective effort of the World Bank, European Bank for Reconstruction and Development, and the European Investment Bank covering regions of Europe, North Africa, Central Asia, and the Middle East.

The aim of our research is to pinpoint the most significant determinants of total annual sales for an enterprise in Poland using the data compiled in the enterprise surveys. Thanks to random forests, we identified features corresponding with labor force and government adherence as most important. Our goal is to quantify differences between enterprises with differing amounts of full-time employees, amount of government adherence to regulations and labor capital thanks to federal grants. Our target variable is d2, In last fiscal year, what were the establishment's total annual sales, and we will focus strictly on enterprise performance in Poland for the year 2019 since the data for Poland is available only for this year.

Hypothesis: Enterprises that cooperate and adhere to government regulations receive more labor capital grants for full-time permanent employees translating to additional sales.

Random forests are a flexible and effective machine learning algorithm for identifying variable importance of our regression task based on Gini statistic of the individual trees. When splitting a node, it searches for the best feature amongst a random subset of features resulting in wider diversity and a stronger model. As will be further discussed later, random forests were vital for our acquisition of most significant features in a dataset full of missing values.

2. Dataset and Variable Description

The Enterprise Surveys were composed using a global methodology including standardized survey instruments and uniform sampling of manufacturing and service sectors in regions of the world. The survey consists of two stages, a phone screening to determine eligibility of the establishment and a two-version questionnaire for either manufacturing or service industries. The goal of the surveys is to study establishment performance which conduct business in a physical location, to qualify for the survey, the establishment is required to have its own management and control over its workforce.

Our initial variable selection method to analyze growth of total annual sales was to focus on explanatory variables centered around investing. We naturally inferred that companies which

invest in training, assets, R&D and human capital are more likely to grow and succeed. Unfortunately, as seen in the table 2.1 below, our explanatory variables were plagued by an extremely high amount of missingness thus we needed to change our approach.

Table 1. Initial variable selection with high missingness

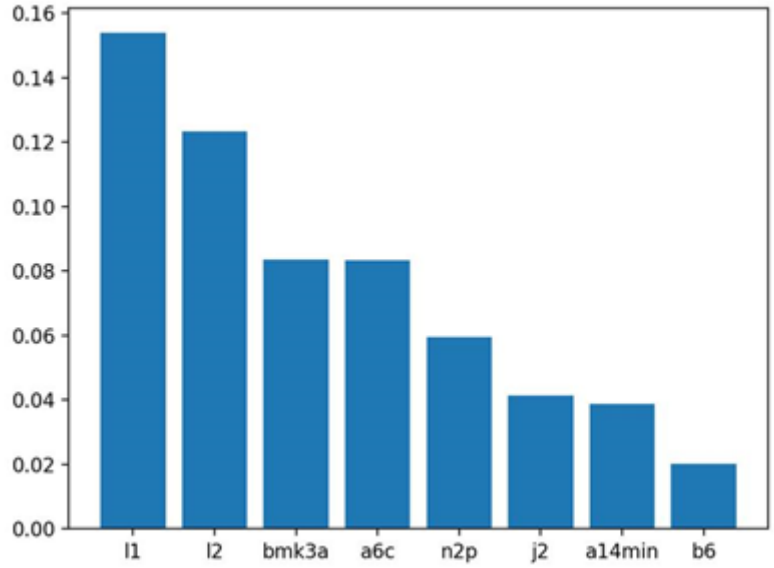
Variable	Percent Missing
d2	46.6
h9	96.42
l1	4.09
l11a1	96.57
n2a	67.57
n5a	75.97
n5b	74.73
n5c	89.92

The below table represents our final selection of important variables chosen to perform our analysis on. Our variable selection method involved using all variables from the entire ES dataset and re-coding all "-9" values for "NaN". Any variables with missingness greater than 30% were filtered out. Consequently, multiple imputation was conducted on the missing values and a random forest was performed on the filtered variables using the built-in module of scikit-learn package in Python. The results of the random forest generated the most important variables pictured in table below in order of importance.

Table 2. Description of selected variables

Variable	Type	Importance	Description
d2	double	Target	In last fiscal year, what were the establishment's total annual sales?
l1	long	0.153888	Number of permanent, full-time employees at the end of last fiscal year
l2	long	0.123431	Number of permanent, full-time employees at the end of 3 fiscal years ago
bmk3a	byte	0.083668	Percent of working capital in government grants
a6c	int	0.083406	Screener size
n2p	double	0.0596	Total cost of sales in last fiscal year
j2	byte	0.041506	What percentage of senior management time was spent in dealing with government regulations
a14min	byte	0.038716	Minutes
b6	int	0.020148	Number of full-time employees of the establishment when it started operations

Graph 1. Variable importance after running a Random Forest model



We would like to now present our rational backed by literature on why we agree with the final variables produced by the random forests in explaining growth of total annual sales. Half of our variables (b6, bmk3a, l2 and l1) correspond with labour capital in an enterprise. According to Caroline Freund, global director for trade, investment and competitiveness of the World Bank Group, large firms with greater numbers of employees will innovate and achieve higher productivity, they also grow markets and create demand in their supply lines thus contributing to sales (Ciani, 2020). Regarding variable b6, one of the most critical factors influencing business growth is beginning size of a firm regarding full-time permanent employees (Davidsson, 2002). Lastly for variables j2 and bmk3a, the role of enterprises cooperating closely with government procedures and regulations opens opportunities for federal subsidies. In a study done on the role of government subsidies on a firm's productivity, enterprises that received grants up to a critical level experienced significantly higher productivity (Qun, 2012). Negative impacts on productivity came to light when excess government assistance was allotted (Qun, 2012).

3. EDA, analysis of dataset and variables

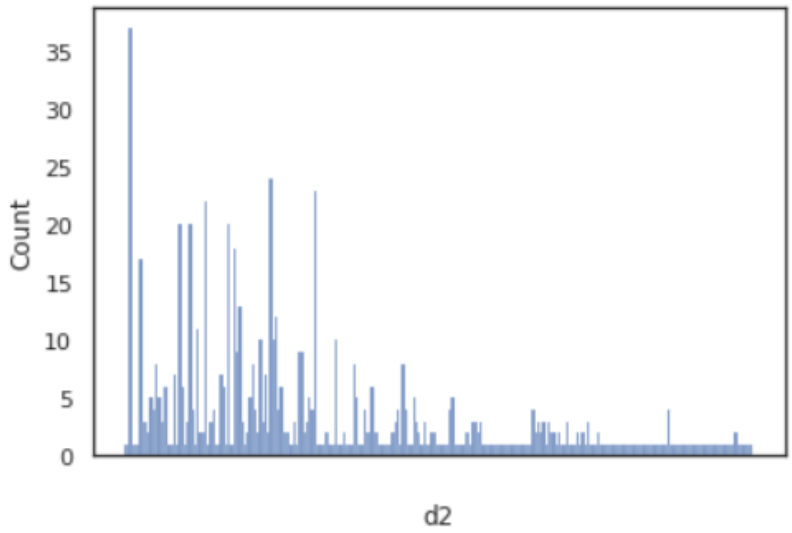
This section presents the process of variable exploration and selection for the final model. In our analysis we decided to measure overall sales in the company in the last fiscal year. In this case, d2 is our target variable. We produced some visualisations to better show the relationships between variables in the dataset.

Table 3. Missing value count for each variable before pre-processing.

Variable	Number of missing observations
d2	638
l1	56
l2	308
bmk3a	290
a6c	0
n2p	1206
j2	358
a14min	0
b6	177

The above table shows how many missing values are present in our dataset. Our target variable has around 50% of missing values. As it is the most important variable in the dataset, we decided to delete all observations with missing values for this variable. After deleting missing data from d2 it turned out that n2p still had a majority of missing data, so it also got deleted. The remaining variables were all numerical and had around 20% of missing data so we considered replacing missing values with mean values for each variable.

Graph 2. Distribution of the target variable d2



Graph above shows the distribution of our decision variable - Total sales in the last fiscal year. We can see that the distribution has no pattern except for the fact that there are more companies

with lower total sales than companies with highest total sales, which is to be expected because small and medium-sized companies outnumber larger companies. The sales were very varied, starting from $60.000 up to $2.700.000.000.

Table 4. Descriptive statistics of the variables.

Variable	Minimum	Maximum	Mean	Standard Deviation
d2	60000	2700000000	30637325.7	138403431.9
l1	1	3500	68	201.9
l2	1	3500	63.4	211.4
bmk3a	0	30	0.2	1.9
a6c	1	3500	74.4	230.6
j2	0	100	10.2	17.7
b6	1	2550	42.7	194.6

Variable	Lower_3_sigma_rule	Upper_3_sigma_rule
d2	-384572970.075	445847621.498
l1	-537.786	673.979
l2	-571.008	697.892
bmk3a	-5.760	6.183
a6c	-617.481	766.370
j2	-43.056	63.564
b6	-541.201	626.776

Table 4. Count of missing data of selected variables and 3 sigma rule calculation

These 2 graphs show some basic statistics related to the variables in the dataset. On the first sight we can see that our decision variable has big differences between minimum and maximum values, which is also visible in Table 2. To deal with missing data we decided to use the 3-sigma rule. The lower bound didn't give us too much insight because all values were below the minimum values for our variables but the upper bound showed that in order to deal with the missing data using this rule, we would have to delete most of the observations. At the end we decided against deleting any more of the observations and instead we replaced them with mean values for each column.

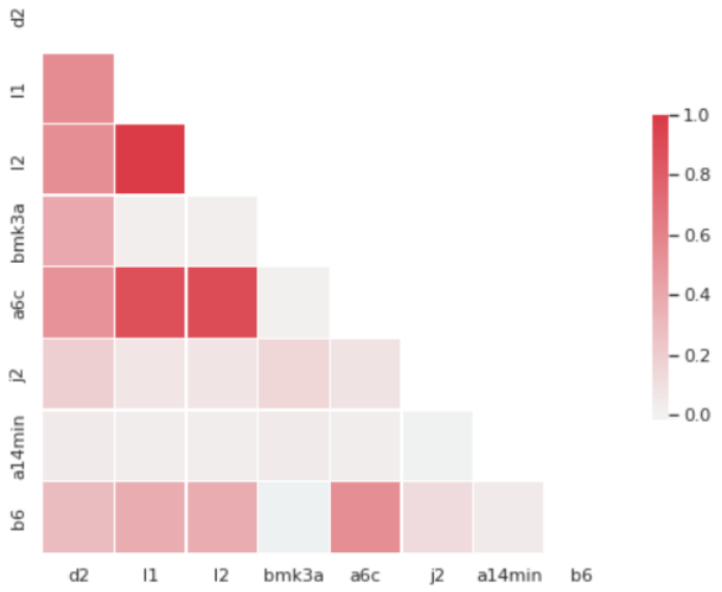


Table 4. Pearson correlation heatmap

Table 4 shows a heatmap of correlations between variables in the dataset. In general, this dataset is quite correlated. Variables l1, l2 and a6c have a very high positive correlation of around 90%. It seems like these 3 variables are mostly correlated between themselves. The reason for it could be that all 3 of them refer to the size of the company. L1 and l2 store information about the number of employees and a6c about its screener size. Another important fact is that our decision variable has a medium correlation of around 50% with most other variables, except for a14min, which shows that we should expect biased results.

4. Predicting sales before imputation

A linear regression can be calculated in R with the command lm. For this multiple regression model, we will regress the dependent variable, d2, on all predictor variables:

$$d2 = \beta_0 + \beta_1\, l1 + \beta_2\, l2 + \beta_3\, bmk3a + \beta_4\, a6c + \beta_5\, n2p + \beta_6\, j2 + \beta_7\, b6$$

We decided to use MLR because we have cross-sectional. Consequently, no mixed model for longitudinal data was required. We will illustrate the result of multiple regression and demonstrate the importance and difference of inspecting, checking, and verifying your missing data before accepting the results of your analysis.

For calculations we used the dataset with 1,369 observations. Therefore, on the table below shows that the model reduced number of observations to 97. The R-squared is 0.974, meaning that approximately 97% of the variability of d2 is accounted for by the variables in the model. In this case, the adjusted R-squared indicates that about 97% of the variability of d2 is

accounted for by the model, even after considering the number of predictor variables in the model. The coefficients for each of the variables indicates the amount of change one could expect in Total Annual Sales (d2) given a one-unit change in the value of that variable.

The number of permanent, full-time employees at the end of 3 fiscal years ago (l2) is not significant (p=.61107) and the coefficient is negative which would indicate that larger number of permanent employees would lower the Total Annual Sales. Alternatively, Number of permanent full-time employees in the last FY (l1) is significant (p=.03677) but surprisingly the effect is also negative, meaning that for a one unit increase in l1, we would expect a 545 thousand polish zlotys decrease in Total Annual Sales.

Furthermore, Screener Size (a6c) is significant (p=.001), and we would expect sales to increase by 937 thousand polish zlotys per Screener Size unit increase. Total costs of sales (n2p) and number of full-time employees of the establishment when it started operations (b6) are the most significant with an increase per unit of 1 Zloty and a decrease of 237 thousand polish zlotys, respectively.

	Dependent variable:
	Sales
Number Permanent Full-Time Employees last FY	-545,201.900**
	(257,144.400)
Number Permanent Full-Time Employees 3 FY ago	-123,844.300
	(242,662.300)
% of Working Capital in Government grants	58,849.680
	(739,493.500)
Screener Size	937,001.600***
	(275,885.600)
Total Cost of Sales In Last FY	1.137***
	(0.036)
Senior Management % Time Spent In Dealing With Govt Regulations?	-97,513.960
	(124,488.200)
Number Full-Time Employees when the company started operations	-237,547.300***
	(34,037.160)
Constant	1,171,393.000
	(2,256,875.000)
Observations	97
R^2	0.974
Adjusted R^2	0.972
Residual Std. Error	18,388,019.000 (df = 89)
F Statistic	477.969*** (df = 7; 89)
Note:	*p<0.1; **p<0.05; ***p<0.0

5. Missing Data Mechanism and Multiple imputation

5.1 Missing Data Mechanism

5.1.1 Table of missing pattern

a6c	n2p	l1	b6	l2	bmk3a	j2	d2	total
1	1	1	1	1	1	1	1	0
0	0	56	177	250	290	358	638	1769

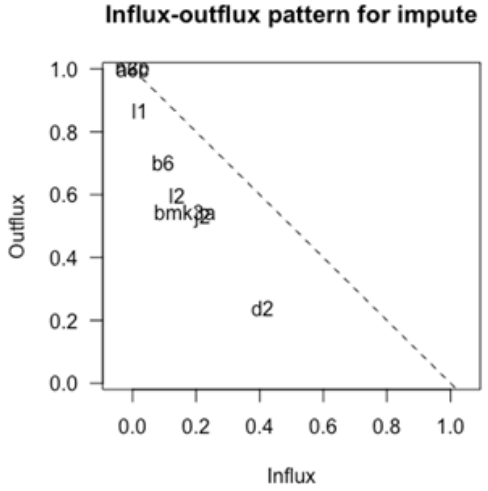
49 missing patterns in total, where bmk3a, j2 and d2 missing the most.

5.1.1 Influx and Outflux for all variables.

The function flux() in mice calculates Ij and Oj for all variables. The value of Ij depends on the proportion of missing data of the variable. Influx of a completely observed variable is equal to 0, whereas for completely missing variables we have Ij = 1. For two variables with the same proportion of missing data, the variable with higher influx is better connected to the observed data and might thus be easier to impute.

Variable	pobs	influx	outflux	ainb	aout	fico
d2	0.533966	0.409234	0.239118	0.841469	0.082666	0.406293
l1	0.959094	0.023522	0.868853	0.55102	0.167229	0.669459
l2	0.817385	0.140041	0.596382	0.734857	0.134687	0.612154
bmk3a	0.788167	0.164761	0.54381	0.74532	0.127367	0.597776
a6c	1	0	1	0	0.184598	0.68298
n2p	1	0	1	0	0.184598	0.68298
j2	0.738495	0.220625	0.526286	0.80846	0.131553	0.570722
b6	0.870709	0.096483	0.700396	0.715093	0.14849	0.635906

From above we can see influx of n2p and a6c is 0 and outflux of n2p and a6c is 1.



From the plot above we can see a6c and n2p located higher up together in the display as they are more complete with 0 missing values and thus potentially more useful for imputation.

5.2 Missing data mechanism check

5.2.1. MCAR Test

Little´s MCAR test

statistic	df	p.value	missing.patterns
377	238	2.18E-08	49

The null hypothesis of the MCAR test is that the data is missing completely at random. Since its p-value is 0.0000000218 given high statistic value and low p-value, our data is not missing completely at random.

5.3 Imputation using the mice package

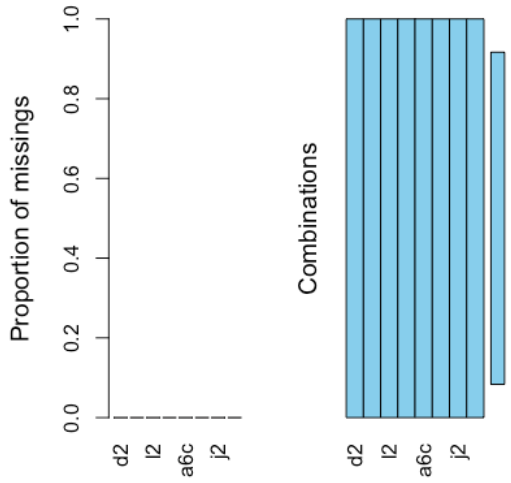
The package creates multiple imputations for missing data. The MICE algorithm can impute mixes of continuous, binary, unordered categorical and ordered categorical data.

data = impute, method = “cart” (Classification and regression trees) m = 1 (Number of multiple imputations) maxit = 20 (number of iterations) seed = 500 (setting seed for reproducibility)

```
TempData <- mice(impute, method = "cart",maxit = 20, m = 1,seed = 500)
```

5.3.1 Confirming the imputation

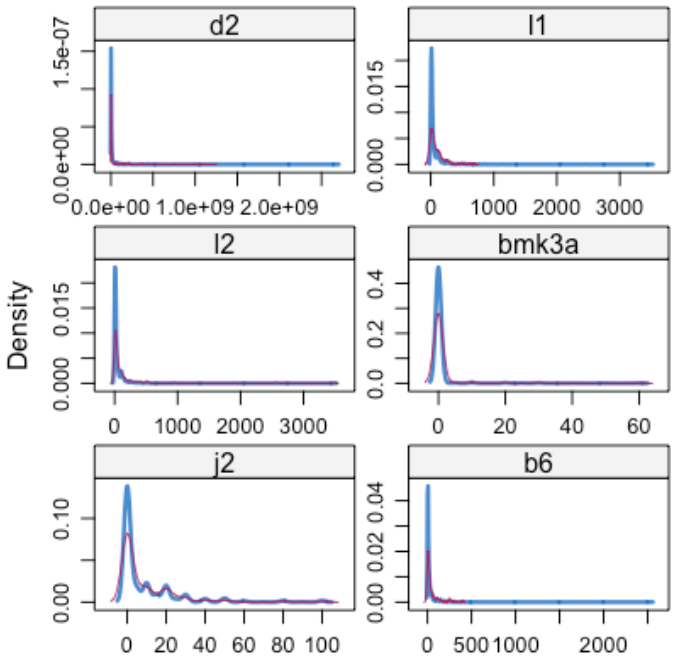
Confirming if there is no missing data after imputation.



From the visualization above we can see there is no more data missing in any of the variables hence all the missing values were replaced with imputed values

5.3.2 Density Plot

The density plot produces the figure by which shows kernel density estimates of the imputed and observed data. In this case, the distributions are very similar and match very well.



6. Predicting sales after multiple imputation

After replacing the missing values in the previous section by applying the multiple imputation method, we ensure that the replaced missing values do not impose an enormous bias when modelling the present relationship between variables. The idea behind of multiple imputation is to use the distribution of the observed variables to estimate different possible values for the missing observations. Under certain conditions we can ensure that these estimations are not biased since we can infer the true value and get the standard errors around the estimations many examples can be also found in the book of van Buuren (2018).

Missing data occurs practically in also every research, regardless on the type, control, and design of study. The problem with missing data is that it drastically reduces the predictive power of estimates and therefore produces biased estimates, thus, leading to wrong conclusions and recommendations.

In many cases, when missing data is present, a common practise (but wrong when not considering the type of missing data) is to drop the missing observations, however this is at cost of statistical power in the model. This practice is valid when the missing value is not missing at random (MNAR) meaning that missing values depend on the hypothetical value, or

it is dependent on another variable's value. However, some parametric models like the ordinary least square (OLS) model cannot deal with missing data and some imputation method must be done in order to ensure predictive power in our model. The idea is to keep the sample size as large as possible and avoid dropping the rows with missing observations that decreases our sample size. Note that the larger the sample size, the closer we get to the true values, and therefore we reduce bias in the model.

Going back to our sales estimation example, in the section regarding the linear regression model previous imputation, we can observe that model was trained only with 97 observations, which is quite low to rule out variables considering the sample size of the initial dataset. Therefore, we can be certainly sure that the model will have a considerable improvement after imputation, and we can see the changes in the following table regarding the new estimates and confidence intervals.

	Dependent variable:
	Sales
Number Permanent Full-Time Employees last FY	297,597.200***
	(40,041.490)
Number Permanent Full-Time Employees 3 FY ago	-39,700.190
	(28,600.620)
% of Working Capital in Government grants	3,060,505.000***
	(677,199.300)
Screener Size	68,111.180**
	(34,575.370)
Total Cost of Sales In Last FY	229,173.600*
	(117,846.500)
Senior Management % Time Spent In Dealing With Govt Regulations?	567,889.600***
	(170,003.300)
Number Full-Time Employees when the company started operations	83,460.170***
	(22,909.790)
Constant	-5,353,270.000
	(3,325,027.000)
Observations	1,369
R^2	0.291
Adjusted R^2	0.287
Residual Std. Error	99,340,855.000 (df = 1361)
F Statistic	79.727*** (df = 7; 1361)
Note:	*p<0.1; **p<0.05; ***p<0.01

In the above table we can appreciate that although our sample size increased in 1369 observations, the adjusted R^2 has decreased considerably. This decrease is attributed to 2 factors:

Firstly, the initial model without imputed values was biased because of having a small sample size. Secondly, the correlations between the features in our model (explained in section 3) is quite high, which means that we have a problem of multicollinearity which also impacts on our estimates and bias them by increasing the variance, making the model less robust. Since our model has a problem of multicollinearity it is not the best model to explain the variance in Sales, however, this model is certainly less biased than the first one.

In this sense we can attempt to interpret the above model. First, we see that all variables are statistically significant with 95% confidence. Except for Number Permanent Full-Time Employees 3 FY ago which is highly correlated with the Number Permanent Full-Time Employees last FY variable.

We can see that having a new permanent employee increases sales by approximately 298 thousand polish zlotys. Also, 1 % increase of working capital in government grants increases in average sales by roughly 3 million polish zlotys.

The variable screener size refers to the number of workers in the establishment as determined by the screener questionnaire. Therefore, having 1 employer more increases sales by 68.1 thousand polish zlotys in average, note that it is less than those permanent workers from the previous fiscal year, which is in accordance with the economic law of diminishing marginal returns, meaning that adding an additional factor of production (labour) results in smaller increases in output (sales).

Regarding our costs of sales variable, by increasing 1 polish zloty, sales increase by 230 thousand polish zlotys in average. This variable is linearly related to sales, therefore, if sales increase, the cost will do so, and vice versa.

Also, 1 percentage increase in the time that senior managers spend in dealing with governmental regulations increases productivity and therefore increases sales by nearly 570 thousand polish zlotys in average. Lastly, the number of full-time employees when the company started the operations (initial size of the company) determines the potential growth of the company and therefore for 1 extra employer at the beginning, the company can get in average an increase in sales by roughly 83 thousand polish zloty in average per year.

Conclusions

This cross-sectional study covered 1369 companies in Poland for the year 2019, and our analysis was fully focused on the factors impacting sales. We bring an alternative way to select the most impactful variables on sales based on a Random Forest algorithm in Python. Also, we have imputed our dataset using the famous R package MICE, under the same framework of R we ran all regression models. One of the biggest limitations in this study was the amount of missing data present in the variables which makes it harder to calculate unbiased estimates. We overcame this limitation by using the multiple imputation method so that we can ensure a large sample size and run our regression model and get less biased estimates. We can highlight that there's a huge difference in the model coefficients when running a model before imputation (few observations) with biased estimates, and another model with imputed missing values. The predictive power completely changes and value of the coefficients. This study can be improved by doing a further analysis on validating all OLS assumptions, also by dropping correlated variables and solving the issue of multicollinearity presented in the last regression model after imputing the values. To summarise, the number of permanent employees, their productivity and the amount of grants received are the factors that positively impact on sales the most.

References

Andrea Ciani, Marie Hyland, Nona Karalashvili, Jennifer L. Keller, Alexandros Ragoussis, Trang Thu Tran (2020). Making It Big: Why Developing Countries Need More Large Firms, World Bank Group, DOI: 10.1596/978-1-4648-1557-7

Per Davidsson, Bruce Kirchhoff, Abdulnasser Hatemi–j & Helena Gustavsson (2002) Empirical Analysis of Business Growth Factors Using Swedish Data, Journal of Small Business Management, 40:4, 332-349, DOI: 10.1111/1540-627X.00061

Van Buuren, Stef. 2018. Flexible imputation of missing data. Chapman and Hall/CRC. Available online: https://stefvanbuuren.name/fimd/

Qun, B. (2012). Government Subsidies and Firm's Productivity——An Empirical Study Based on Chinese Industrial Plants. China Industrial Economy.

Codes

Python code: Variable selection with Ranfom Forest

```
# load dataset

df = pd.read_csv('./dataset_project.csv',
low_memory=False)
df = df[df.country == 'Poland']
df['a14y'].unique()
df.reset_index(drop=True, inplace=True)


# convert "-9" into NaN

for col in df.columns:
    df[col] = df[col].replace(to_replace=-9,
value=np.nan)

# set target variable

y = df.d2

# filter variables with <30% of missing data

filter = df.isnull().sum()/len(df) < 0.3
df = df.loc[: , filter]
df = pd.concat([y, df], axis=1)

# label encoding (note: this is different
than one-hot encoding)

numerical =
df.select_dtypes(exclude=['object'])
categorical =
df.select_dtypes(include=['object'])

list = []
for i in categorical.columns:
    df[i] = df[i].astype('category')
    df[i] = df[i].cat.codes
    list.append(df[i])


columns = categorical.columns
categorical =
pd.DataFrame(np.array(list).transpose(),
columns=columns)


df = pd.concat([numerical, categorical],
axis=1)
sorted = df.isnull().sum()/len(df)
sorted.sort_values()

# Apply multiple imputation to the values

imp_mean = IterativeImputer(random_state=0)
imp_mean = imp_mean.fit(df[df.columns])
df[df.columns] =
imp_mean.transform(df[df.columns])

sorted = df.isnull().sum()/len(df)
sorted.sort_values()


# target and features selection

X = df.loc[ : , df.columns != 'd2']
y = df['d2']

# split dataset
np.random.seed(seed = 42)
X_train, X_valid, y_train, y_valid =
train_test_split(X, y, test_size = 0.2,
random_state = 42)


# train random forest
rf = RandomForestRegressor(n_estimators =
100, n_jobs = -1, oob_score = True, bootstrap
= True, random_state = 42)

rf.fit(X_train, y_train)

# get variable importance
var_importance = dict(variables =
X_train.columns, importance =
rf.feature_importances_)
var_importance = pd.DataFrame(var_importance)
var_importance.sort_values(by=['importance'],
inplace=True, ascending=False)

# variables that affect to "d2" by more than
2%
top_variables =
var_importance.query('importance > 0.02')
plt.bar(top_variables['variables'],
top_variables['importance'])
plt.show()
```

R code: Imputation

```
library(mice)

#Loading the dataset

data <- read.csv("./dataset.csv")
#filtering dataset for selected variable
impute=df[c("d2",
"l1","l2","bmk3a","a6c","n2p","j2","b6")]
tempData <- mice(impute,
                 method = "cart",
                 maxit = 20,
                 m = 1,
                 seed = 500)
data <- read.csv("./dataset.csv")
#filtering dataset for selected variable
impute=df[c("d2",
            "l1",
            "l2",
            "bmk3a",
            "a6c",
            "n2p",
            "j2",
            "b6")]

tempData <- mice(impute,
                 method = "cart",
                 maxit = 20,
                 m = 1,
                 seed = 500)
```