USING MACHINE LEARNING TO BUILD AN OPTIMAL MODEL THAT FORECAST THE NET EXPORT OF CEREALS IN ROMANIA

Robert Ştefan **SBÎRCEA**[1]**,** George Marian **CĂLIN**[2]**,** Iulia Bianca **BOGOS**[3]
[1]PhD Student, Bucharest University of Economic Studies
Email: robertsbr21@yahoo.ro
[2]PhD Student, Bucharest University of Economic Studies
Email: calingeorge18@yahoo.com
[3]PhD Student, Bucharest University of Economic Studies
Email: iulia_bogos@yahoo.com

Abstract

The objectives of this article are to build forecasting models that estimate the net exports of several kinds of cereal in Romania. A comparison between the traditional model of regression (the linear model) and machine learning algorithms will be made. Moreover, there will be built an ensemble of these models in order to minimize the RMSE. The implications are complex and represent a mixture of mathematical, econometric and programming skills. The results of the article aimed to identify new methods that increase the confidence level of estimates.

Keywords: *Machine Learning, Linear Models, Ensemble, Forecast, RMSE, Random Forest, Extreme Gradient Boosting*
DOI: 10.24818/CAFEE/2020/9/11

Introduction

Romania is famous for "the granary of Europe". But how much of this statement is true and how much has become a myth? From year to year, farmers' problems increase: lack of irrigation, bad weather, difficult to access funds and subsidies. Even so, agriculture is one of the most important pillars of the Romanian economy, contributing significantly annually to the form of Gross Domestic Product.

A traditional corn exporter, Romania has managed to maintain its position as a market leader in 2019, with a quantity of over 3.6 million tons. Practically, 73% of all the corn sold by Europe over in third countries came from Romania. The second exporter was Bulgaria, with over 1 million tonnes, followed by France with around 113,000 tonnes. Wheat also remained in second place after France, which sold about 12.9 million tonnes. However, the quantity has increased significantly, from about 3.7 million tons to 5.3 million tons this year. Compared to the entire quantity of wheat sent by Europe across borders, Romania's market share was 15.2%.

Trade does have a short-term impact on aggregate demand. One element of aggregate demand is net exports; a change in net exports alters the aggregate demand curve and has a short-term impact on real GDP. A decrease in net exports reduces aggregate demand, whereas a rise in net exports increases it, all else being equal.

The main determinants of net exports are domestic and foreign income, relative price levels, exchange rates, domestic and foreign trade policies, and tastes and techniques. If the current account is positive, spending on the purchase of goods and services will outweight the outflow, and the economy will be in the black (that is, net exports will be positive in a simplified analysis). If the current account is negative, spending on goods and services leaving the country exceeds spending on entering the country, and the economy runs a current account deficit (that is, net exports are negative in a simplified analysis).

In order to support the rise of GDP, the level of net export has to be estimated. As the technology is developing, our aim is to find a modern alternative to the classic models. To do so, we are going to build several models that will overcome this challenge. Our intention is to develop the analysis of the available data by executing a fully supervised machine learning workflow.

1. Literature Review

Adler J. (2009) provided great examples for statistical computing and data visualization in his book and also quoted the Founder of Kaggle, who claimed that R is a fast becoming lingua franca of statistics. Brett L. (2019) stated that the core of Machine Learning is transformation of information into actionable intelligence and without Machine Learning it will be almost impossible to keep up with the enormous stream of information. Pierobon G. (2018) described in his article a complete Machine Learning workflow.

He presented a comprehensive checklist that people should follow in supervised learning, both for regressions and classifications. Paruchuri V. (2012) proved that ensemble learning can outperform any other single model and ensembles can improve performance by adding other models to the ensemble. Rajbangshi A. (2020) presented the similarities and differences between the two types of ensemble methods (bagging and boosting). Both methods are better in performance compared to single models. Therefore, boosting method reduces both bias and variance, while bagging method helps reduce only the variance. Kaushik S. (2017) ended his article by claiming that ensembleing is an effective and popular technique that is frequently used for beating the accuracy benchmark of the best individual algorithm.

In their study, Akalpler E.& Shamadeen B. (2017) found that exports and economic growth are positively correlated in the United States, and the positive impact of gross fixed capital creation is a major contributor to this relationship. However, high import levels and unemployment are impeding economic expansion. They recommended several measures that can help achieving bigger economic growth. Additionally, steps can be taken to boost export levels even more.

The employment of export subsidies and incentives can achieve this. Trade agreements may be created to remove trade restrictions. Bilateral and international trade agreements may also strengthen this. It is necessary to reduce the threat that imports are posing. The link between imports and economic growth that is unfavourable can be attributed to inflation's consequences. Therefore, precautions must be taken to guarantee that imports do not fuel inflationary pressure. This could be accomplished by levying high taxes on consumer goods and offering enticements to import producer items.

What’s more, Al Hemzawi B. & Umutoni N. (2021), support the above articles and claims that there is a strong direct correlation between the net export and the GDP. On the other hand, the Import-led Growth hypothesis is supported by the positive and statistically significant association between imports and GDP. In light of these findings, it is suggested

that the government enhance its import regulations by encouraging non-direct imports of consumables used in the production of commodities with a surplus that can be exported. Therefore, similarly Romania had an import-led Growth, that needs to be optimized, in order to maximize general outlook.
Next, it will presented a modern methodology that will estimate the net export and overcome the disadvantages of traditional models.

2. Setting and Exploratory Data Analysis

It will be made an accurate estimation of the performance of net export of the following cereals: wheat, corn, barley and rice. Numeric outcomes have to be predicted, therefore this is a multivariate supervised machine learning problem, thus regression techniques will be used. Our data is in "csv" format. It presents a header row with the column names. It seems to contain both: categorical and numeric information.
The dataset has 59 rows and 12 columns, one of which is our response/target variable. Next, a glimpse will be taken at the data and it will be observed just the first few rows of the data frame:

Table 1: Visualization of the data set

Observations	59						
Variables	12						
Year	2007	2008	2009	2010	2011	2013	…
Culture	Wheat	Wheat	Wheat	Wheat	Wheat	Wheat	…
Area (1000 ha)	1975	2110	2149	2162	1947	2104	…
Average production kg/ha	1541	3403	2421	2688	3663	3468	…
Total production (1000 tons)	3044	7181	5202	5812	7132	7296	…
UM	Lei/kg	Lei/kg	Lei/kg	Lei/kg	Lei/kg	Lei/kg	…
Average Price	0.61	0.66	0.46	0.59	0.88	0.85	…
Cant_Imp (tons)	587526	441637	628843	719954	559139	679828	…
Val_IMP (1000 EUR)	108969	93295	77874	112444	124236	122897	…
Cant_Exp (tons)	206634	1988758	2340673	2480143	1568734	4773294	…
Val_Exp (1000 EUR)	45937	381635	302947	379446	309769	977680	…
Exp_Net (1000 EUR)	-63032	288340	225073	267002	185533	854782	…

Source: Own processing based on the data set from https://www.madr.ro/culturi-de-camp/cereale.html

The response variable is `Exp_Net (thousand euro)` and it has the following statistics:

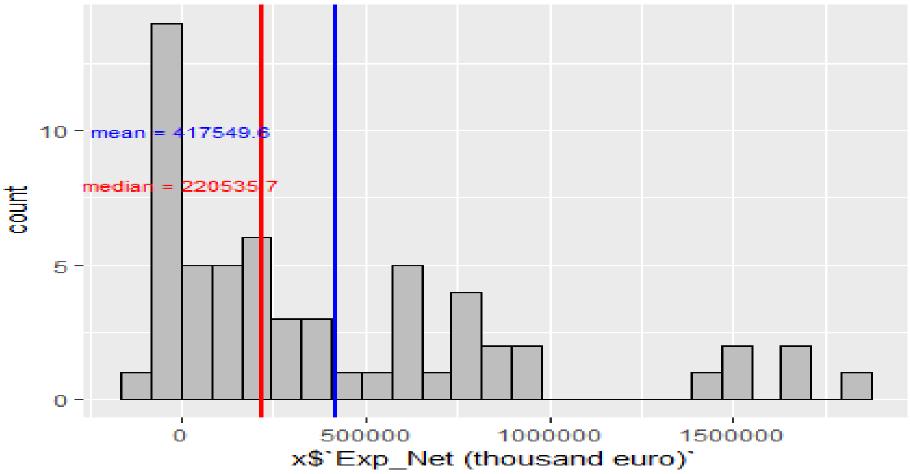
Table 2: Statistics of the response variable

Min.	1st Quantile	Median	Mean	3rd Quantile	Max
-114090	1138	215998	417416	653810	1842061

Source: Own processing based on the response variable

The net export ranges from -114090 to 1842061. The median and the mean are not close, but since the median is actually smaller, this results in a strong skew of the variable distribution to the right.

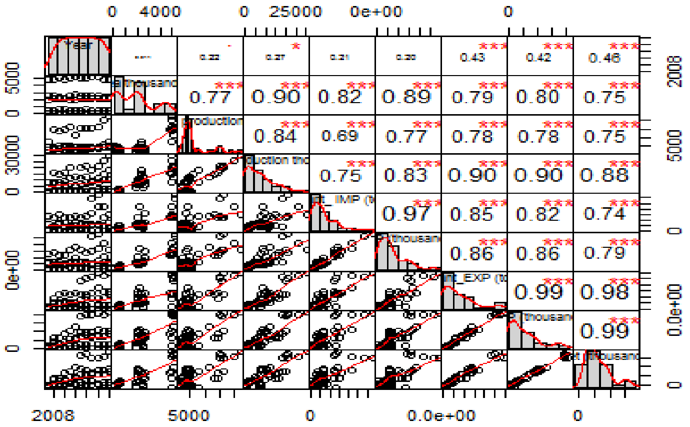
Figure 1. The histogram of the response variable



Source: Own processing based on the response variable

The analysis will be followed by a correlation plot. The correlation between all variables will be presented in a chart. This is the moment to evaluate if all the variables need to be part of the model.

Figure 2. The correlation plot between all variables



Source: Own processing based on the data set

After analysing all these covariates, we are going to select only 4 for in our equation: Area, Average production kg/ha, Cant_Imp and Cant_Exp, the last two being the amount of product imported and exported. We've already seen in the correlation plot presented before that there seems to be a significant correlation between some features. We want to make sure that multicollinearity is not an issue that prevents us to move forward. To do this, we will compute a score called Variance Inflation Factor (VIF) which measures how much the variance of a regression coefficient is inflated due to multicollinearity in the model. If the VIF score is more than 10, multicollinearity is strongly suggested and we should try to get rid of the features that are causing it.

Table 3. Variance Inflation Factor of the covariates

Covariate	**Area thousand ha**	**Average production Kg/ha**	**Cant_IMP (to)**	**Cant_EXP (to)**
VIF	8.960997	6.833532	8.645124	8.531014

Source: Own processing based on the linear model

None of the covariates surpasses the threshold of 10 so we will consider multicollinearity not to be a big issue. However, some would argue that it could indeed be a problem to have as many features with scores of 7. The actual data set has limitations from the point of view of the number of observations. We have 60 samples, so this is definitely a small dataset. We will divide the dataset into train and test sets and make sure we use cross-validation when we train our model. In that way, we ensure we are using our few observations as well as we can.

3. Modeling

The process of modeling will be based on the following assumptions:

- A list of 5 models (Linear Model, Support Vector Machine with the radial kernel, Random Forest, XGB Tree – extreme gradient boosting tree, XGB Linear) is going to be trained at the same time, allowing the use of 5 fold cross-validation for each model.
- Furthermore, parallel processing will be used to boost speed and the scope of the paper will not focus on the nature of algorithms. Results will be commented on and interpreted by comparing the performance over training and test sets, focusing on RMSE (root mean squared error) as our metric.
- In addition, an ultimate combination of models will be build by ensembling the model list and then stacking them in order to improve performance even more.

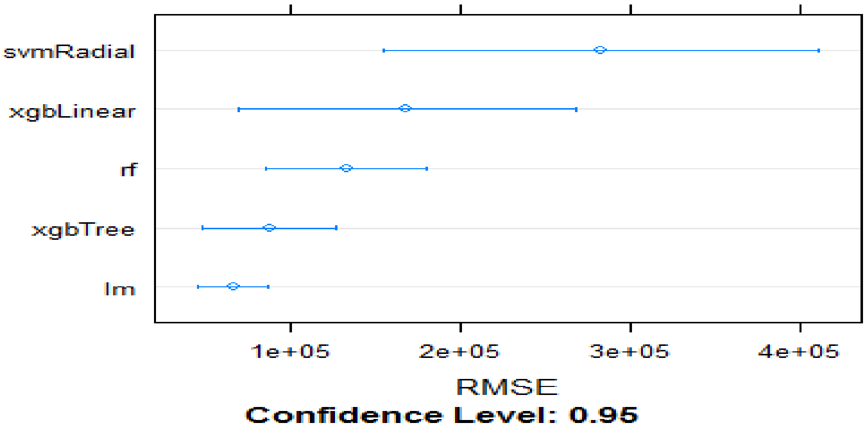
After training the above mentioned models we are obtaining the followings RMSE:

Table 4. RMSE of the 5 models

Model	LM	SVM	RF	XGBT	XBTL
RMSE	66281	282808	132984	87634	168890

Source: Own processing based on the trained models

Figure 3. The confidence levels of RMSE allocated to the 5 models



Source: Own processing based on the trained models

The Linear Model has the lowest RMSE and it looks like the strongest candidate for the final model. Actually, we need to take a look at how models cope with predictions. But first, let's analyse the correlations between all these models:

Table 5. Correlation between the trained models

	lm	svmRadial	rf	xgbTree	xgbLinear
lm	1.00000	-0.0521	0.9230	0.00937	-0.2747
svmRadial	-0.05212	1.0000	-0.1883	-0.97855	0.2104
rf	0.92301	-0.1883	1.0000	0.17374	0.0467
xgbTree	0.00937	-0.9785	0.1737	1.00000	-0.1654
xgbLinear	-0.27473	0.2104	0.0467	-0.16538	1.0000

Source: Own processing

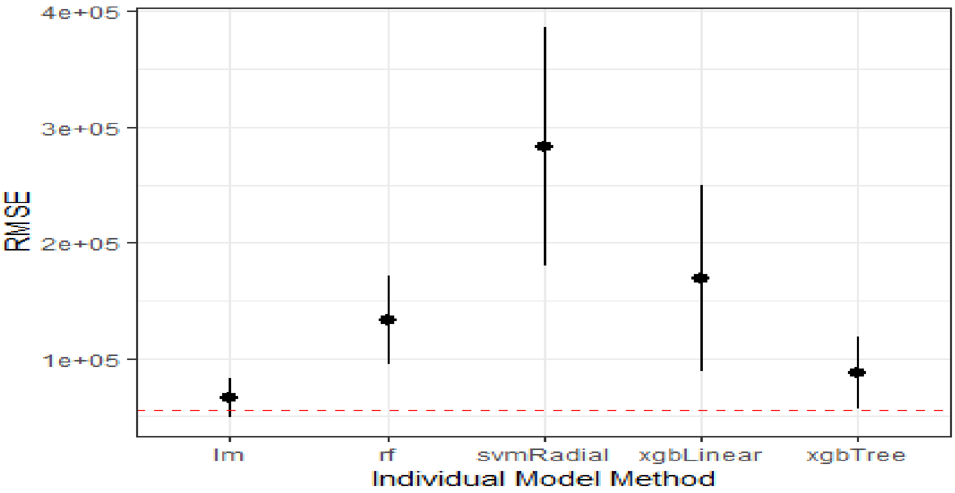
The Linear Model has a strong correlation with random forest, while svmRadial has a strong correlation with xgbTree. The only model that has a low correlation with the others is xgbLinear. The next step is to ideally ensemble models with low correlations and there will be created a new model by combining our 5 models to find the best possible model, that optimizes performance. Thus, a linear combination of the models will be performed, reducing the RMSE value to 55226 and the following weights were returned:

Table 6. Ensemble's weights

Model	LM	SVM	RF	XGBT	XBTL
Weights	0.7582	-0.0047	-0.2792	0.3646	0.1618

Source: Own processing

Figure 4. RMSE of the ensemble



Source: Own processing

The ensemble's RMSE performance is represented by the red dashed line, where it is clear that the ensemble outperforms the other models. In addition to this, another ensemble is going to be built, using a generalized linear model via penalized maximum likelihood with a gaussian distribution as the stacking method. Therefore, in this case, the RMSE was reduced to 54008, where the optimum parameters were alpha = 0.01 and lambda = 901.

Table 7. Optimum parameters for the ensemble

alpha	lambda	RMSE	Rsquared	MAE
0.10	901	54008	0.986	40886
0.10	9013	56138	0.984	42149
0.10	90130	82544	0.974	61466
0.55	901	56218	0.984	42927
0.55	9013	54901	0.984	41734
0.55	90130	89934	0.980	71266
1.00	901	56042	0.984	41962
1.00	9013	54572	0.985	40785
1.00	90130	105778	0.985	84639

Source: Own processing

Finally, our models have to run over unseen data (test set) in order to evaluate the performance. To do so, the test set is going to be predicted with each model.

Table 8. Observed values vs predictions

Original	pred_lm	pred_svm	pred_rf	pred_xgbT	pred_xgbL	predict_ens1	predict_ens2
799947	855367	600575	808685	709708	853061	803678	801121
922183	1022576	701227	806242	804168	844818	963941	955851
599778	663647	605425	574701	591901	631463	649253	645166
748444	851250	581208	713785	576250	844967	777336	766789
148750	152659	185830	184804	135427	137113	136144	138133
189437	182201	161381	168160	176730	126134	176812	177485
-14825	-30553	4190	-13859	-19219	-12473	-24033	-24618
-18003	-35718	18672	-24769	-28855	-29088	-30840	-31508
1533651	1642700	571410	1229089	1104388	1342963	1496770	1477793
1842061	2021889	569165	1234783	1059688	1342934	1766399	1730962

Source: Own processing

The next step is to compute the RMSE:

Table 9. RMSE of every model

Model	ensemble_1	ensemble_2	LM	SVM	RF	XGBT	XGBL
RMSE	35627	44191	85239	516392	218790	291313	175752

Source: Own processing

It is no wonder that by using ensembles the performance got better and so ensemble_1 outperforms every other model. Even though ensemble_1 has the lowest RMSE and it is the most optimal model, let's have a look at the variation of each model.

Table 10. The variation of observed vs predicted values of every model

pred_lm	pred_svm	pred_rf	pred_xgbT	pred_xgbL	predict_ens1	predict_ens2	Closest model
0.0693	-0.2492	0.0109	-0.1128	0.0664	0.00466	0.00147	predict_ens2
0.1089	-0.2396	-0.1257	-0.128	-0.0839	0.04528	0.03651	predict_ens2
0.1065	0.00941	-0.0418	-0.0131	0.0528	0.08249	0.07567	pred_svm
0.1374	-0.2234	-0.0463	-0.2301	0.129	0.0386	0.02451	predict_ens2
0.0263	0.24928	0.2424	-0.0896	-0.0782	-0.08475	-0.07138	pred_lm
-0.0382	-0.1481	-0.1123	-0.0671	-0.3342	-0.06664	-0.06309	pred_lm
1.0609	-1.2826	-0.0652	0.2964	-0.1587	0.62109	0.66057	pred_rf
0.9841	-2.0372	0.3759	0.6028	0.6157	0.71309	0.75018	pred_rf
0.0711	-0.6274	-0.1986	-0.2799	-0.1243	-0.02405	-0.03642	predict_ens1
0.0976	-0.691	-0.3297	-0.4247	-0.271	-0.04107	-0.06031	predict_ens1

Source: Own processing

The linear model has 2 of the closest predictions, the random forest has 2 of the closest predictions, SVM has one of the closest prediction, ensemble_1 has 2 of the closest prediction, while the ensemble_2 has 3 of the closest prediction.

Table 11. The test set's response variable by Year and Culture

Year	Culture	Exp_Net
2017	Wheat	799947
2018	Wheat	922183
2017	Corn	599778
2018	Corn	748444
2017	Barley	148750
2018	Barley	189437
2017	Rice	-14825
2018	Rice	-18003
2017	All	1533651
2018	All	1842061

Source: Own processing

The table above represents the test set (Exp_Net) and there were added other columns to interpret the results. For the net export of wheat, the best model was ensemble_2, for corn there was no best model, because each year had a different model being the best, for barley the linear model was the best, while for rice the random forest model was the most optimum. On the other hand, from a macroeconomic perspective, the thing that matters is the net export of all cereals, and here the best model was ensemble_1, with a variation of -2% in 2017 and -4% in 2018.

Conclusions

To conclude, it was observed that the capacity of the Romanian market to export cereals is capped, no matter how much production was gained, because it seems that it is difficult to make collaborations with other counterparties. In 2017, the total production raised by 25%, while in 2018 it raised by 16%, but the exported quantity of cereals was almost constant. Therefore, the errors from above suggest that the model returns a remarkable forecast even though the price per tone was not known, neither for imports nor exports.

Under the current circumstances, the weather has a high weight when the net export is estimated. Thus, the model can be improved by adding this new covariate in order to lower even more the errors of the predictions. In addition to this, the proposed model can be used at a macroeconomic level and the output can serve as an input in next years' budget or in the forecast of GDP.

In order to achieve sustainable economic development, the government should aim to improve market access for exports of goods and services, foster firm capacity to enter and grow in the export market, and establish an export growth facility. To improve the quality of its exports and the amount of money it can earn, the Romanian government has to make technological advancements that can aid in processing its main export commodities in order to improve the GDP growth.

References

1. Rittenberg L., Tregarthen T. (2017). Principles of Macroeconomics, Version 3, Chapter 15. Net Exports and International Finance
2. Adler, J. (2009). R in a Nutshell
3. Brett, L. (2019). Machine Learning with R - Third Edition
4. Kaushik, S., (2017). How to build Ensemble Models in machine learning? (with code in R)
5. Paruchuri, V. (2012). An Intro to Ensemble Learning in R
6. Pierobon, G. (2018). A comprehensive Machine Learning workflow with multiple modelling using caret and caretEnsemble in R
7. Rajbangshi, A. (2020). Types of Ensemble methods in Machine learning
8. Akalpler, E., Shamadeen, B. (2017). The role of net export on economic growth in United States of America. Journal of Applied Economic Sciences, Volume XII, Summer 3(49): 772– 781.
9. Al Hemzawi B. & Umutoni Nn (2021). Impact of exports and imports on the economic growth.