

CA 2 MSc Data Analytics

An analysis of Bovine Meat Production in Ireland

22 May 2022

James Garza

SBA1905

Table of Contents

[Table of Figures 1](#_Toc104149485)

[Abstract 2](#_Toc104149486)

[Introduction 3](#_Toc104149487)

[Data Understanding 3](#_Toc104149488)

[Plots 4](#_Toc104149489)

[Statistics 8](#_Toc104149490)

[Sentiment Analysis 13](#_Toc104149491)

[Data Preparation 13](#_Toc104149492)

[Modelling 14](#_Toc104149493)

[XGBoost 15](#_Toc104149494)

[SVR 15](#_Toc104149495)

[Bayesian Ridge Regression 16](#_Toc104149496)

[Grid Search CV 16](#_Toc104149497)

[Evaluation 17](#_Toc104149498)

[Dashboard 18](#_Toc104149499)

[Conclusion 18](#_Toc104149500)

[Access to Github Link 18](#_Toc104149501)

[References 19](#_Toc104149502)

# Table of Figures

[Figure 1. Effects of EU and USA Subsidies 2](#_Toc104143949)

[Figure 2: Substitution Effect 2](#_Toc104143950)

[Figure 3: Bovine Growth Rate 4](#_Toc104143951)

[Figure 4: Month to Month Percentage Change Line Plot 4](#_Toc104143952)

[Figure 5: Month to Month Percentage Change Box Plot 5](#_Toc104143953)

[Figure 6: Population Growth 5](#_Toc104143954)

[Figure 7: GDP Growth 6](#_Toc104143955)

[Figure 8: GDP Growth per Capita 6](#_Toc104143956)

[Figure 9: Trade Surplus 7](#_Toc104143957)

[Figure 10: Carcass Yields 7](#_Toc104143958)

[Figure 11: Irish Bovine Production QQ Plot 8](#_Toc104143959)

[Figure 12: Ireland Germany Bland-Altman Plot 9](#_Toc104143960)

[Figure 13: Ireland USA Bland-Altman Plot 10](#_Toc104143961)

[Figure 14: Ireland(X) and Germany(Y) Shift Plot 11](#_Toc104143962)

[Figure 15: Ireland(X) and USA(Y) Shift Plot 11](#_Toc104143963)

[Figure 16: Mann-Whitney Results 12](#_Toc104143964)

[Figure 17: Consumer and Producer Sentiment Vs Steak Prices 13](#_Toc104143965)

[Figure 18: PCA Number of Components 14](#_Toc104143966)

[Figure 19: KMeans Elbow Method 15](#_Toc104143967)

[Figure 20: Time Series Cross Validation Split 16](#_Toc104143968)

[Figure 21: Model Results Time Series Cross Validation 17](#_Toc104143969)

[Figure 22: Model Results Standard Cross Validation 17](#_Toc104143970)

# Abstract

Irish beef is one of the top when it comes to quality. One of the main reasons for this is that Irish beef is grass fed verified by Bord Bia (“Irish Beef”). Production in Ireland of bovine meat has increased over 59%, meanwhile population growth has only been just over 30%, these numbers are based on data from the CRO and will be plotted later in this report. With the increase one could conclude that Irish beef is mainly exported, and this will be explored as well. What is contributing to this exponential growth? And can it be predicted? One factor could be the European Union’s Common Agriculture policy enacted in 1962, almost a third of the annual EU budget goes to agriculture (European Commission). Bovine meat is only a part of EU agriculture and part of the EU budget in subsidies. Subsidies have created inefficiencies in the agriculture sector, the EU and USA have the highest farming subsidies in the world. These subsidies are making harder for farmers to compete in global markets. The EU CAP is looking to move away from subsidies over the next few years. Subsidies is an indirect tax to consumers and distorts prices (Parkin et al., Pgs. 140, 160-161, and 464). See figure 1 below from page 161.

Diagram

Description automatically generated

Figure 1. Effects of EU and USA Subsidies

Additionally, to the price controls economic conditions can affect consumer behaviours. There are two phenomena in economics the substitute and wealth effect. When the economy is strong and assets people own, they tend to spend more than they normally would. Equally when the economy is in a recession the substitute effect is where people are not feeling as wealthy and will substitute premium items for lower cost items (Cfa Institute, pg. 337) (Parkin et al., pg. 184-185). Figure below from page 184:

Diagram

Description automatically generated

Figure 2: Substitution Effect

Considering the above economic theories, the data brought into the research will test these theories. Machine Learning will be used to attempt to predict output based on GDP per capita, bovine substitutes and trade.

# Introduction

Do people eat more bovine meat with an increase of wealth? Can production of bovine be predicted? Data will be collected to answer these questions. The datasets collected are monthly and therefore the scope of this research will be on monthly data from Ireland, Germany and USA. Ireland and Germany production will be compared to see if there is any effect from the EU Common Agricultural Policy, and then compared to US production, though both countries highly subsidise agriculture. The machine learning will predict the production of beef in Ireland in metric tonne. The features or inputs will be month of year, GDP, GDP per capita, yield of carcass, population, export/import/surplus, beef prices (in USD), pork and lamb production. These inputs are following in line with wealth and substitution effects, there could be other factors affecting production such as price of feed, petrol prices but this is outside the scope of this research. Using these inputs to predict production could help farmers and factories to produce in line with consumer demands. In tough times there could be bovine meat demand and bovine meat substitutes such as pork or lamb would be in higher demand. Equally in good times there could be a higher demand for bovine meat and less for pork or lamb. Going through this research has led me to learn more about farming than I ever wanted to know. The CRISP-DM project management methodology was used for this research.

# Data Understanding

The data collected from different countries that use different weights and currencies adjustments must be done to be able to compare properly. For example, in the US weight used is pounds and in the EU kilos. Something new that I learnt was the spelling of tonne and ton have two different weight measures. To keep the comparison, equal this research will use weight of meat productions rather than heads as animal size could differ but weights are consistent. Collecting Irish bovine production was a bit of a challenge. Monthly data only goes as far back as 1992 in certain categories and to bring all datasets this will be the start date of the research. Additionally using an end date to ensure all datasets are of the same time length for comparison. Collecting datasets from many sources helped to verify what data is available for analysis. For Irish and German bovine meat production data was collected from EuroStat. The USA bovine meat production was collected from the United States Department of Agriculture Economic Research Service. Population data and global beef prices were collected from United States Federal Reserve Economic Data. Finally, the trade and carcass yield data were collected from the Food and Agriculture Organization of the United Nations. The Food and Agriculture Organization from the UN has excellent data.

Once data had been collected US bovine meat weight is converted into metric tonnes for comparison. There were additional adjustments such as changing US hogs to EU pork. The final adjustment was to the price of bovine meat from Federal Reserve as it is in cents per pound, and this was converted to dollars per kilo.

## Plots

In the following plots the colours are assigned as follows, green for Ireland, gold for Germany and blue for USA. These colours follow colours in the country’s flags. Bovine meat production is much higher in US as compared to Ireland and Germany and this is due to population. Ireland has the highest growth rate of 59.66% and Germany is in negative growth of -38.43%.

Chart, line chart

Description automatically generated

Figure : Bovine Growth Rate

The month-to-month percentage change Ireland must have quite a few changes in its bovine meat production shown in the following plots:

Chart

Description automatically generated

Figure : Month to Month Percentage Change Line Plot

Chart, box and whisker chart

Description automatically generated

Figure : Month to Month Percentage Change Box Plot

Population growth is the fastest growing in Ireland followed by USA and Germany, 40.36%, 28.45% and 3.24% respectively.

Chart, line chart

Description automatically generated

Figure : Population Growth

GDP growth as a percentage was fastest in Ireland followed by USA and Germany, 362.55%, 90.30%, and 36.93% respectively. The 2008 financial crises drop are clearly seen in the plot.

Chart, line chart

Description automatically generated

Figure : GDP Growth

GDP growth per capita was in the same order as GDP which is a good sign that economic growth is affecting all the population. Ireland is again on top with 362.55% growth, USA at 148.66% and Germany at 91.82%.

Chart, line chart

Description automatically generated

Figure : GDP Growth per Capita

Looking at trade subtracting imports from exports and creating a trade surplus Ireland has the highest followed by Germany and finally USA.

Chart, line chart

Description automatically generated

Figure : Trade Surplus

Finally, Ireland has the lowest carcass yield and USA the highest, followed by Germany. Ireland could improve its exports by not producing more bovine meat but by increasing carcass yields.

Chart, line chart

Description automatically generated

Figure : Carcass Yields

## Statistics

Statistical analysis of the datasets using Ireland as a baseline was first to test if the distributions of the dataset are normal. See visualization of the QQ-plot for Irish bovine meat production. The QQ-plot which orders the z-scores from low to high and plots the values (Bruce et al., pg. 71). Seen in plot below that the distribution is not normal.

Chart, line chart

Description automatically generated

Figure : Irish Bovine Production QQ Plot

To confirm the visual results of the normal distribution was the Shapiro Wilk test which returned a pValue of 0.00218 (“7.2.1.3. Anderson-Darling and Shapiro-Wilk Tests”).

Comparing Ireland to Germany visually to see if the datasets have agreement or not Ireland and Germany have better agreement than Ireland and USA. This could be due to EU Common Agricultural Policy. The tests used were the blandaltman which are testing means of two datasets and finding agreement or lack of it (Bland and Altman).

Graphical user interface, scatter chart

Description automatically generated

Figure : Ireland Germany Bland-Altman Plot

Chart, scatter chart

Description automatically generated

Figure : Ireland USA Bland-Altman Plot

Visualizing using a shift plot shows how much different the two bovine meat productions are. Using Pingouin’s visual library of the plot shift has some very good visualizations. Ireland is X and Germany is Y values in figure 14.

Diagram

Description automatically generated

Figure : Ireland(X) and Germany(Y) Shift Plot

Background pattern

Description automatically generated with medium confidence

Figure : Ireland(X) and USA(Y) Shift Plot

Now to compare the countries for significance in Ireland vs Germany and Ireland vs USA. Since the distributions of the datasets are not normal the statistical test will be the Mann-Whitney. This is used when comparing two non-normal independent simple random samples. Mann-Whitney compares the medians of the dataset (Weiss, pg. 493-497). This non-parametric test gave back the following results.

Graphical user interface, text

Description automatically generated

Figure : Mann-Whitney Results

Based on those P-value being less than alpha (0.05) the null hypothesis are rejected on both comparisons as there is statistical significance between datasets.

## Sentiment Analysis

Web scraping was a new experience and trying to think of the best way to approach this was a challenge. Twitter is not a simple solution and not sure if people are genuine on twitter to gauge sentiment. The Irish Farmers Association website would represent a more accurate sentiment of farmers or producers. Search terms used for sentiment were beef, cattle, and prices this produced 444 articles from the IFA. Getting consumer sentiment from the Irish Examiner with the same search terms produced over 1,220 articles. The python library used was Selenium and made the web scraping task simpler. Now that articles are collected it is on to analysing the sentiment using the NLTK library. The text is cleaned removing any unneeded syntax or words. The next step was to get Irish beef prices and I went with the price of steaks and got this data from the CRO. The article dataset needed to be resampled to monthly and getting the mean of the daily sentiment. This was then compared to the prices to see if there was any correlation which there was negative correlation. See plot below.

Timeline

Description automatically generated with medium confidence

Figure : Consumer and Producer Sentiment Vs Steak Prices

The dataset is small regarding months, but the consumer and producer sentiment are positively correlated, and beef prices do not seem to be affected. This could be due to price controls in Ireland.

# Data Preparation

Combing all the collected data and putting into on dataset to prepare for machine learning. Some feature engineering is a good way to help the machine learning algorithm get higher prediction results in detecting patterns. (Müller and Guido, pg. 224) (Auffarth, pg. 68) (Jansen, pg. 12) (Aurélien Géron, pg. 27). Features added were a 12-month moving average, 12 months high and low and month of the year.

The dataset is quite small and since the statistical method of cross validation grid search will be used this is a powerful and stable alternative to train, test split (Muller et al, pg. 258) (Aurélien Géron, pg. 77). Brining in the month required for it to be one hot encoded and I used various scalers, standard, robust and min max to see which is best. Using sckits learn column transformer allowed these transformations to happen within pipeline. Scikit Learn Pipeline function is a way to streamline the machine learning process. This was the first time to use this function and initially there were very low predictions from the machine learning models. Upon further investigation the scaler was put in the wrong area of the process and was scaling the one hot encoded data. This led to further analysis of what is happening under the hood to ensure that all the transformations are happening in the right sequence. Pipeline is an effective tool to streamline and automate the iterative process of machine learning on new data. I do not foresee ever doing machine learning without pipeline again.

In the data outliers could affect the outcome of the models but should not be discarded but experimented with what data does the model perform the best as removing them without experimentation could be a mistake. The approach taken here is to remove them via the interquartile range. The formula is as follows (Ali et al., pg. 3):

Interquartile range (Bruce et al., pg. 38):

Models will be trained with and without outliers and see if there is an improvement in the performance of the models and will be recorded for comparison.

# Modelling

Model selection is the next step after the data has been prepared. Testing various models in this research to see which fits best to this dataset. The three implemented machine learning algorithms are XGBoost, Support Vector Regressor and Bayesian Ridge with polynomial features. Pipeline allowed to mix unsupervised models with supervised and looking to add improvement to model performance. PCA and is used to reduce dimensionality or noise in a dataset. This dataset had a lower number of dimensions but using pipeline made so easy that it had to be implemented. With PCA in number of components using the 0.95 allows PCA to pick the dimensions that capture 95% of the variance (Vanderplas, pgs. 439-445). This would seem easier to implement in pipeline than using the elbow method.

Chart

Description automatically generated

Figure : PCA Number of Components

KMeans is another unsupervised machine learning algorithm and finds clusters in datasets. The goal of clustering is to split the data to where the data points are quite similar. Using the elbow method to find the number of clusters is components and found on this dataset 5 is the best number of components for clustering (Müller and Guido, pg.168) (Aurélien Géron, pg. 240). These methods reduce dimensions and possibly noise from the dataset and possibly help add speed to the models training (Aurélien Géron, pg. 215).

Chart, line chart

Description automatically generated

Figure : KMeans Elbow Method

## XGBoost

XGBoost is an optimized implementation of gradient boosting that can provide quite accurate results on many types of datasets (Aurélien Géron, pg. 210). Implementing this ensemble algorithm which uses weak learners into strong learners (Wade, pg. 53) (Liu, pg. 237). This was the first time that I have used XGBoost but it was quite simple to implement and fit well inside the grid search cross validation. The XGBoost performance increased with the additions of PCA.

## SVR

Support Vector Machine using the regression version which usually finds the largest margin between vectors in regression it tries to put as many vectors as possible in the margin (Aurélien Géron, pg. 164). Additionally, this model does a polynomial regression without having to use the polynomial function (Müller and Guido, pg. 229). Finding the best parameters using Grid Search and implanting dimensionality reduction.

## Bayesian Ridge Regression

Bayesian Ridge Regression is a probabilistic model that estimates and is of the regression model family as such (“1.1. Linear Models — Scikit-Learn 0.24.1 Documentation”). The formula is as follows:

Bayesian Ridge does work well with time series as most linear regression models do (Downey, pg. 243). Using Bayesian in conjunction with Polynomial features can help find coefficients within the features and improve the model’s performance. The implementation of this was not so simple as the hyper tuning of not just the Bayesian model but also the polynomial features.

## Grid Search CV

All the above model parameters can then be run through and tested for best performance. The Sklearn library of Grid Search CV runs through all the parameters in a grid like pattern. This can take lots of time to do but since there were not many parameters and the dataset is small this was the best search algorithm from SKlearn (“Sklearn.model\_selection.RandomizedSearchCV — Scikit-Learn 0.21.3 Documentation”). Since this is a time series it is important that the time element remains intact, otherwise you would lose the sequence of events. In the cross-validation section of the GridSearchCV the cross validation implemented is the SKLearn time series split that works well with time series. This research has done split with and without time series splits and see which performs best. The important part is not to shuffle the observations see image below from ‘Practical Time Series Analysis’ (Nielsen, pgs. 343-346).

Chart

Description automatically generated

Figure : Time Series Cross Validation Split

# Evaluation

In this research using a dictionary to store the model names and model function made life simple. Additionally storing the hyper tuning parameters was easier as well, combining this with pipeline the algorithm was to loop through, models, scalers, outliers or not and capture all the results within 1/3 of the code done in the last CA. Using a class object of machine learning allowed for the dataset to be changed and information to be stored and returned in an efficient way. The output of each trained model is captured and stored for later analysis. The metrics used to capture the results is the R2, Means Squared Error and Mean Absolute Error. Finally, regarding using cross validation as mentioned before it is not recommended to use the standard cross validation but to use a time series split to avoid any shuffling of the dataset (Nielsen, pgs. 343-346). Here is the comparison of the two results with R2 score above 50% average performance per model.

Chart, bar chart

Description automatically generated

Figure : Model Results Time Series Cross Validation

Chart, bar chart

Description automatically generated

Figure : Model Results Standard Cross Validation

The results are not much difference on average of around 5% but would be worth it being sure that the model is general enough and works well in production.

## Dashboard

Creating a dashboard was no easy feat but using Dash really helped. Initially I was going to use explainer dashboard, but it could be overwhelming with all the options for the user. A similar approach is warranted. The only dashboard this research was able to produce was the dash with drop down showing the different machine model predictions. This research wanted to have additional options in the dashboard such as allowing the user to put in elements to see a prediction and then displaying a probability of that prediction. Additionally, for the users to see the results of the sentiment analysis. Great plans but ran out of time.

# Conclusion

This research was able to show that there is a correlation between bovine meat and economic conditions. The inputs month of year, GDP, GDP per capita, yield of carcass, population, export/import/surplus, beef prices (in USD), pork and lamb production were able to produce good predictions. The state of the economy effects consumers behaviours and with the wealth and substitution effect this allows for good predictors for demand for bovine meat products.

# Access to Github Link

<https://github.com/jamesgarza73/DataAnalyticsLevel9>

# References

“1.1. Linear Models — Scikit-Learn 0.24.1 Documentation.” Scikit-Learn.org, scikit-learn.org/stable/modules/linear\_model.html#bayesian-ridge-regression. Accessed 22 May 2022.

“7.2.1.3. Anderson-Darling and Shapiro-Wilk Tests.” Www.itl.nist.gov, www.itl.nist.gov/div898/handbook/prc/section2/prc213.htm. Accessed 22 May 2022.

Agresti, Alan, and Maria Kateri. Foundations of Statistics for Data Scientists : With R and Python. Boca Raton, Crc Press, 2021.

Aurélien Géron. Hands-on Machine Learning with Scikit-Learn and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems. O’Reilly Media, Inc., Sept. 2019.

Bland, J. Martin, and Douglas G. Altman. “Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement.” International Journal of Nursing Studies, vol. 47, no. 8, Aug. 2010, pp. 931–936, 10.1016/j.ijnurstu.2009.10.001.

Bruce, Peter C, et al. Practical Statistics for Data Scientists : 50+ Essential Concepts Using R and Python. Sebastopol, Ca, O’reilly Media, Inc, 2020.

Cfa Institute. Economics Level 1 CFA. New York ; Munich, Learning Solutions, 2011.

Downey, Allen. THINK BAYES : Bayesian Statistics in Python. S.L., O’reilly Media, 2021.

European Commission. “The Common Agricultural Policy at a Glance.” European Commission, 2019, ec.europa.eu/info/food-farming-fisheries/key-policies/common-agricultural-policy/cap-glance\_en. Accessed 21 May 2022.

“Irish Beef.” Irish Food & Drink, www.irishfoodanddrink.com/irish-beef/. Accessed 18 May 2022.

Kirk, Matthew. Thoughtful Machine Learning with Python : A Test-Driven Approach. Sebastopol, Ca, O’reilly, 2017.

Liu, Sebastian. MACHINE LEARNING with PYTORCH and SCIKIT-LEARN : Develop Machine Learning and Deep Learning... Models with Scikit-Learn and Pytorch. S.L., Packt Publishing Limited, 2022.

Müller, Andreas C, and Sarah Guido. Introduction to Machine Learning with Python : A Guide for Data Scientists. Beijing, O’reilly, 2017.

Nielsen, Aileen. Practical Time Series Analysis : Prediction with Statistics and Machine Learning. Sebastopol, Ca, O’reilly Media, Inc, 2019.

Parkin, Michael, et al. Economics. Harlow, England Pearson, 2017.

Vanderplas, Jacob T. Python Data Science Handbook : Essential Tools for Working with Data. Beijing Etc., O’reilly, Cop, 2017.

Wade, Corey. Hands-on Gradient Boosting with XGBoost and Scikit-Learn : Perform Accessible Machine Learning and Extreme Gradient Boosting with Python. Birmingham, Packt Publishing, 2020.

Weiss. Introductory Statistics. Harlow, Pearson, 2015.

Wallpapertops.com, 2015, wallpapertops.com/walldb/original/d/b/a/707646.jpg. Accessed 18 May 2022.

Europa.eu, 2022, ec.europa.eu/eurostat/estat-navtree-portlet-prod/BulkDownloadListing?file=data/apro\_mt\_pwgtm.tsv.gz.