An Evaluation of Machine Learning Approaches for Milk Volume Prediction in Ireland

Christian O'Leary
Nimbus Research Centre
Munster Technological University
Cork, Ireland
christian.oleary@mtu.ie

Conor Lynch
Nimbus Research Centre
Munster Technological University
Cork, Ireland
conor.lynch @mtu.ie

Abstract—Milk yield production strongly influences energy consumption, plant utilisation and farm revenue. The competency to identify the annual peak week and the ability forecast daily, weekly, or annual lactation curves in advance is beneficial across the agri-business chain at a management, processor, and farm gate level. The value of a milk yield prediction system depends upon how accurately it can predict inherent milking patterns and its ability to adjust to factors affecting milk volumes.

This study presents a review of machine learning (ML) approaches applied to dairy-specific data and meteorological signals to predict milk volumes in Ireland at a national level. This contrasts with existing approaches that forecast for individual cow or herd levels in Ireland. The resulting model performances serve as a benchmark for any future algorithms developed.

Adopting the Lewis scale, it was shown that Random Forest and K-Neighbours Regression derived "highly accurate forecasts" with an average Mean Absolute Percentage Error of 8.28 and 8.35 respectively for prediction timelines of 1-52 weeks ahead. 13 other models are shown to produce "highly accurate" week-ahead forecasts which degrade over increasing forecast horizons. This includes three existing ML milk yield forecasting algorithms to facilitate a comparison with the existing state-of-the-art.

Keywords—machine learning, lactation curves, milk yield forecasting, open-source, scikit-learn

I. INTRODUCTION

With the abolition of milk quotas in April 2015, EU dairy farmers were free to expand production without purchasing milk quota rights for the first time in over 30 years [1] thereby realising its full potential in terms of output and export earnings [2]. This, with a desire to help meet an expected 20% increase in the global consumption of milk and dairy products by 2050, has culminated in milk production cycles emanating larger peak to trough ratios (PTRs).

During the period of 2010/11 to 2015/16, Ireland saw an increase of milk production by approximately 26%. The PTR has increased from 4.7 to 5.7 during this period and as of 2019 is approaching 10 [3]. Ireland now produces over ten times as much milk in the peak as it does in the trough (May versus January). For comparison, the UK maintains a relatively stable PTR of just 1.2, according to the Agriculture and Horticulture Development Board (AHDB) [4].

This problem is more acute in countries with primarily pasture-based dairy farming systems such as Ireland, where milk production cycles are heavily influenced by the patterns of cattle lactation cycles, unlike countries with stall-based systems [5]. The PTR is further exasperated by farmers adjusting the date of calving to maximise the use of grazed grass in the cows' diet in an attempt to produce milk at as low a cost as possible [3]. Thus, unless diversity from the grass-based dairy farming (which itself is seasonal and conditioned by weather variations) is managed, Ireland will continue to see increasingly pronounced PTRs and supply/demand curves.

The recent Irish Farmers Association (IFA) factsheet [3] on Irish Dairying stated that milk supply expansion has increased by over 300 million litres in 2019, taking the overall production to 7.98 billion litres annually. By the end of 2020, national milk yield had increased by a further 3.8% to 8.29 billion litres [6]. With a consequent increase in herd sizes (averaging 90 cows) and milk production (averaging 450,000 litres), the ability to predict future production levels using traditional empirical curve fitting prediction models developed on lower levels of milk production becomes problematic [5]. Milk yield instability translates to increased risk for all participants in the dairy supply chain. The identification and adoption of a suitable algorithm or methodology to mitigate against this risk will help to ensure that the sector remains competitive and profitable during volatile periods.

In this paper, several novel findings are reported for milk yield forecasting in Ireland. Firstly, multiple present-day machine learning models are implemented including Random Forest, Support Vector Machines (SVM) and K-Neighbours Regression (KNR). Secondly, the models are configured to provide weekly Irish milk yield forecasts at a national level. This differs from existing approaches in literature that have only forecasted localised Irish milk production on individual cow and herd levels. Thirdly, model predictive performance is monitored from 1 to 52 weeks-ahead. Some pre-processing steps are also explored, and their efficacy recorded and explained. Features are engineered and extracted from weather-related signals as well as domain-specific data such as milk volumes and prices. These results are presented along with a review of the current state-of-the-art of milk yield forecasting to provide a comprehensive benchmark of results.

The paper is organised as follows: Section I comments on definitions and motivations. Section II reviews existing forecasting techniques. Sections III and IV detail the benchmark datasets, proposed approach and evaluation metrics used. The results are tabulated and discussed in Section V, while Section VI describes high-level conclusions and recommendations.

978-1-6654-5227-4/22/$31.00 ©2022 IEEE

II. BACKGROUND

When examining milk yield prediction algorithms, a lactation period of 305 days (305-d) is generally considered, as this takes advantage of the 60-day dry period, i.e. the yearly calving interval [7]. Using 4 years of data (2006-2010) from a single research farm in Ireland, Murphy et al. [8] evaluated the ability of a multiple linear regression (MLR) model, a static artificial neural network (SANN), and a nonlinear auto regressive model with exogenous input (NARX) to forecast the total daily herd milk yield from a herd of 140 lactating pasture-based dairy cows over multiple forecast horizons: 10-d, 30-d, 50-d and 305-d. Inspired by biological brain mechanisms, the SANN model described in [8] follows the architecture of a conventional densely connected neural network. The architecture of a SANN consists of multiple layers of neurons: one input layer, one or more hidden layers and one output layer. The SANN used in [8] used one hidden layer with 4 neurons. Each neuron in one layer has a weighted connection to each neuron in the next layer. The inter-layer weights are learned parameters of the model. Each neuron aggregates these weights with a learned bias (used arithmetically as an offset) before passing the result through an activation function. Activation functions are specified as hyperparameters of the model and are not learned automatically. Examples include linear and sigmoid activation functions. The NARX model is a form of recurrent neural network (RNN). RNNs use feedback connections to persist a form of working memory within the model. The NARX model in [8] uses two recurrent connections along with Bayesian regularisation (BR) to fit the model. Each model predicted the daily milk yield levels for a full lactation of the horizon 305-d with a percentage root mean square error (RMSE) of $\leq$ 12.03%. Given the reported effectiveness of these models on Irish milk data, implementations of all three are included in this paper's analysis and are used as an evaluation benchmark for comparison purposes.

Using data from 39 pasture-based Holstein-Friesian (HF) Irish dairy cows, Zhang et al. [9] examined the benefits of utilising weather parameters for milk production forecasting at individual cow level. The NARX model was found to provide a greater prediction accuracy than the MLR model for predicting an individual cow's annual milk yield (kg) with R^2 values greater than 0.7 for 95.5% and 14.7% of total predictions respectively. The superiority of the NARX algorithm is consistent with the findings presented in [8]. From experimental data of 174 lactating HF dairy cows in Indonesia, Soeharsono et al. [10] developed a linear regression equation, obtaining a p-value <0.05, to predict daily milk production (DMP) based on linear body and udder morphometry. Using monthly test day milk yield records for 588 Sahiwal cows over a 49-year period (1961-2009), Gandhi et al. [11], predicted the first lactation 305-d lactation milk yield using an artificial neural network. A total of five monthly test day milk yields (2, 3, 5, 7, 8 monthly test day records) were used in neural networks to train data using the BR algorithm. The best neural network model was shown to predict with 93.18% accuracy. Availing of data supplied by Lely Industries (Maassluis, the Netherlands) for the period Jan. 2015 to April 2016, consisting of 1,094,780 observations of sensor data originated from 57 farms in 6 different countries. Jensen et al. [12] developed tailored dynamic linear model (DLM) designed to forecast milk yields of individual cows per milking. A study by McParland et al. [13], focusing on 24-hour milk yield composition, used a Tru-Test (Auckland, New Zealand) dataset comprising AM and PM part-day milk weights and composition (i.e. fat and protein composition) of 48,737 test days from 23,737 lactations in 17,896 cows recorded between the years 2004 and 2017 on 237 farms. Liseune et al. [14] advocate for the use of deep learning models in daily herd lactation curve modelling using individual cow data rather than aggregated herd averages. Using convolutional neural networks (CNNs) and sequential autoencoders (SAEs), feature extraction and imputation were conducted before feeding the data into a SANN (i.e. a multilayer perceptron) for prediction [14]. Nguyen et al. [15] used animal and dietary features for predicting the individual yield of 36 HF cows over a 10-month period, finding the Random Forest model to be superior in terms of accuracy, although noting that the SVM model is sufficiently accurate while being more computationally tractable.

Due to the main limitation that specific milk yield datasets are highly case specific, most models could be the optimal model based on the specific research objects and test datasets under unique conditions. Therefore, researchers from similar or dissimilar regions do not have a mutual target to compare [5]. Thus, the main objective of this paper is to investigate a range of existing open-source machine learning (ML) algorithms implemented through scikit-learn [16] for national milk yield forecasting in Ireland. This is in contrast to other studies which have focused on smaller scale milk production for individual Irish cows of herds. Due to the relatively small size of the available data (< 700 milk yield instances), this study prioritises traditional ML models such as Support Vector Regression (SVR) and SANNs instead of deep neural networks. These algorithms are then expected to serve as a benchmark for theoretical comparisons, to act as a baseline for emerging state-of-the-art or tailored approaches in Ireland.

III. DATASETS USED

This research used data from several Irish and European sources[1]. Data for the target variable, i.e. milk volume in this instance, is available with a weekly resolution format. Corresponding data includes the number of milk suppliers and grass growth particulars. Publicly available forecast and historical climatic signals comprising temperature and precipitation levels were included through Met Eireann[2] – the Irish Meteorological Service provider. Other examined datasets related to livestock census information, animal feed statistics as well as milk price reports from Ireland and correlating European countries. Milk prices from 27 EU countries were included for feature extraction along with UK prices and EU averages. From the aforementioned input

[1] The publicly available data used in this study are available on request.

[2] https://www.met.ie/

978-1-6654-5227-4/22/$31.00 ©2022 IEEE

channels, appropriate candidate features were extracted to act as model inputs to machine learning forecasting algorithms. Note, not all available features were used; some channels did not produce strong or significant correlations.

Spanning January 2009 to November 2021, the data used included over twelve years of milk volume data. From Figure 1, it can be observed that the milk volumes exhibit a pronounced annual seasonal periodicity.

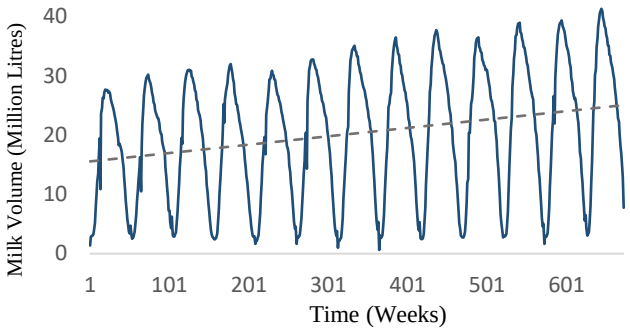

Figure 1: Weekly milk yield in millions of litres (01/2009 – 11/2021)

This regular pattern causes milk yield to endorse a strong autocorrelation function. An excerpt of the prevailing variables displaying high absolute Pearson correlation coefficient values are listed in Table 1. From these, two further variables were engineered to add additional possible predictive value to the available data:

1. The time series week index was mapped to a Cosine wave function where w_i refers to the index of the i^{th} week of the year. This was calculated as per (1):

$$\text{Cos } 2\pi \frac{w_i}{52} \quad (1)$$

   This feature has a strong negative Pearson correlation with the target variable – yielding coefficient of approximately -0.9321.

2. Examining milk yield per supplier, i.e. total weekly milk yield divided by the number of contributing suppliers, this manipulated feature noted a correlation of 0.955 – petitioning it as a prospective model feature.

When considered in conjunction with the original time series channels, further supplementary features can be derived. Due consideration to parameter lags and values at different time periods, e.g. milk yield one week ago (t-1), resulted in additional features of interest being engineered.

Table 1: Strongest variable correlations

Variable	Correlation Strength	Correlation Coefficient
Milk yield	(Autocorrelation)	1.0
Average yield by supplier	Strong Positive	0.95
Week as cosine wave	Strong Negative	-0.92
Number of suppliers	Strong Positive	0.73
Grass growth	Moderate Positive	0.58

The process used to select useful features is outlined as follows: firstly, all weekly values are considered in a batch window of 52 weeks, e.g. milk yield t-1, milk yield t-2, … milk yield t-52, grass growth t-1, … grass growth t-52, and so forth. Historical sampling is conducted for all features except for time-based features. The value of time-based features – e.g. week of year – are instead known at prediction time. These features are all then subjected to a feature scoring process. All features receive a score and are then sorted accordingly in descending order. The scores are plotted as demonstrated in Figure 2. Where the curve is seen to plateau, or elbow, it is assumed that an optimal feature set size has been reached and that adding more features to the set will not result in improved model performance. This approach for feature selection is similar that employed by Gollou et al. [17] and Koprinska et al. [18].

This process was replicated on four occasions using four different feature scoring methods; absolute Pearson correlation, feature importance scores extracted from a fitted Random Forest model, F-regression scores, and Mutual Information scores – resulting in four feature score plots. The F-regression feature score curve is shown in Figure 2 - where the selected feature plateau point is indicated. The feature set sizes for the aforementioned methodologies are presented in Table 2. While this approach is computationally efficient, it suffers from the possibility of human error. The decision on where the curve begins to plateau rests with the data analyst. Consequently, feature sets are therefore susceptible to being potentially either limited or poorly bounded due to human error. This can be partially mitigated by imposing a limit on the number of features shortlisted for display on the plot. A limit of the 100 highest ranking features was found to be conducive to appropriate feature selection, allowing the plot elbow to be clearly visible. For example, the features chosen using Feature Importance scores were: milk yield (t-1), mean supply per supplier (t-1), milk yield (t-2), milk yield (t-52) and week mapped to a cosine wave. All features selected using this method were extracted from lags of the following channels: milk yield, mean supply per supplier, week mapped as a cosine wave, week of year, number of suppliers, and grass growth.

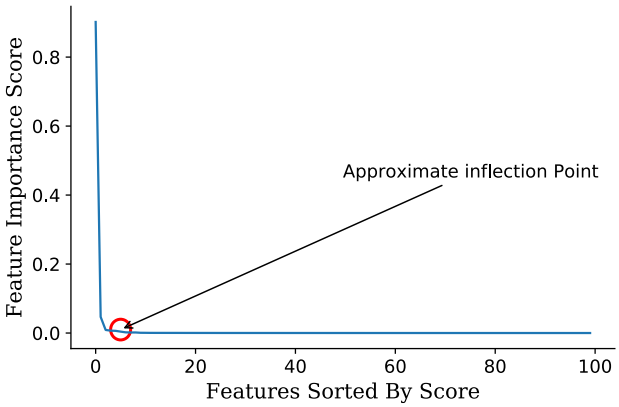

Figure 2: Feature Importance feature scores

Meteorological data was not found to noticeably improve milk yield predictions. This is consistent with the feature score plots and existing literature [9]. Zhang et al. found that any improvements from using weather data were small and may even have been attributable to noise in the data.

978-1-6654-5227-4/22/$31.00 ©2022 IEEE

Table 2: Feature set sizes

Feature Scoring Method	Number Features
Absolute Pearson Correlation	35
Feature Importance	5
F-regression	15
Mutual Information	25

IV. Experiment Design

The milk yield forecasting models from scikit-learn include:

- Bayesian Ridge
- Decision Tree
- Extra Tree
- Gaussian Process
- K-Neighbours Regression (KNR)
- Lasso regression
- MLR
- Linear SVR
- NARX
- Nu SVR
- Passive Aggressive
- Random Forest
- Ridge regression
- SANN
- Support Vector Regression (SVR)

These models are implemented using the Python scikit-learn library[3]. It should be noted that scikit-learn's implementation of Linear Regression can be used as an MLR model when used with multiple input features. Detailed in Table 3, a range of hyperparameter values are specified for each model type. Each model is evaluated using the following procedure:

1. Each dataset was split randomly into training, validation and test sets using an 80-10-10 split*.
2. 50 sets of randomly sampled hyperparameter values were selected for the model.
3. 50 instances of the model were trained on the training data and tested on the **validation set**.
4. The model instance with the lowest validation set error score was selected for testing on the **test set**. This score was recorded as the final model score for that iteration. In performing the model selection before test set testing, leakage is avoided.
5. This process of training, validation and testing is repeated using each of the four feature sets and for all features.
6. The overall process was repeated 30 times to achieve statistical significance. A mean error score was then reported for each model type.

** To ensure models are tested on relevant data, test data was sampled from the post-quota time range, i.e. post April 1st, 2015. Shuffling the data is possible as long as the target variables (y, y+1, y+2, … y+52) are assigned to each observation in advance.*

The SANN and NARX models exhibited poor Mean Absolute Percentage Error (MAPE) scores using the configurations originally outlined in [8]. This can most likely be attributed to the differences in the dataset such as the available set of input features. Consequently, these models were also tuned using the same hyperparameter sampling method as the other models, which resulted in improved MAPE scores. It should be noted that the NARX model makes forecasts recursively, i.e. the model's predictions at step *t* are used as inputs for prediction at step *t*+1. Therefore, data cannot be shuffled in the same manner as with the other models. Instead, the unshuffled data is split chronologically into a training and test set with the test set being the last 10% of observations (which is equivalent to holdout). The model's training loss taken as a proxy for validation score when selecting the best model instance prior to calculating the test score.

Table 3: Model hyperparameters

Model	Hyperparameter	Options
Bayesian Ridge	Num. Iterations	150, 300, 450
	Tolerance	$1e^{-2}$, $1e^{-3}$, $1e^{-4}$
	Alpha 1	$1e^{-5}$, $1e^{-6}$, $1e^{-7}$
	Alpha 2	$1e^{-5}$, $1e^{-6}$, $1e^{-7}$
	Lambda 1	$1e^{-5}$, $1e^{-6}$, $1e^{-7}$
	Lambda 2	$1e^{-5}$, $1e^{-6}$, $1e^{-7}$
Decision Tree	Criterion	MSE, Friedman MSE, MAE
	Splitter	best, random
	Max depth	8, 16, 32, 64, 128, None
	Max features	auto, sqrt, log2
Extra Tree	Criterion	MSE, Friedman MSE, MAE
	Splitter	best, random
	Max depth	8, 16, 32, 64, 128, None
	Max. features	auto, sqrt, log2
Gaussian Process	Alpha	$1e^{-8}$, $1e^{-9}$, $1e^{-10}$, $1e^{-11}$, $1e^{-12}$
	Num. restarts optimizer	0, 1, 2, 3
	Normalize y	True, False
KNR	Num. Neighbors	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
	Weights	Uniform, Distance
	P	2, 3, 4
Lasso	Alpha	0.2, 0.4, 0.6, 0.8, 1.0
	Tolerance	$1e^{-2}$, $1e^{-3}$, $1e^{-4}$
	Selection	Random, Cyclic
MLR	N/A	
Linear SVR	Epsilon	0.0, 0.5, 1.0
	Tolerance	$1e^{-3}$, $1e^{-4}$, $1e^{-5}$
	C	0.01, 0.1, 1.0, 10
	Loss	Epsilon Insensitive, Squared Epsilon Insensitive
	Intercept Scaling	0.001, 0.1, 1.0, 10
	Max. Iterations	500, 1000, 1500
NARX	Hidden Neurons	(2), (3), (4), (2,2), (3,3), (4,4)
	Solver	LBFGS, Adam, SGD
	Alpha	0.00001, 0.0001, 0.001, 0.01, 0.1
	Initial Learning Rate	1, 0.1, 0.01, 0.001
Nu SVR	Nu	0.0, 0.2, 0.4, 0.6, 0.8, 1.0
	C	0.01, 0.1, 1.0, 10
	Kernel	Linear, RBF, Sigmoid
	Degree	2, 3, 4, 5, 6
	Gamma	Scale, Auto
	Coef. 0	0.0, 0.5, 1.0

[3] https://scikit-learn.org/

978-1-6654-5227-4/22/$31.00 ©2022 IEEE

	Shrinking	True, False
	Tolerance	$1e^{-3}$, $1e^{-4}$, $1e^{-5}$
Passive Aggressive	C	0.01, 0.1, 1.0
	Loss	Epsilon Insensitive, Squared Epsilon Insensitive
	Epsilon	0.001, 0.01, 0.1, 1.0
	Average	True, False
	Early Stopping	True, False
Ridge Regression	Alpha	0.2, 0.4, 0.6, 0.8, 1.0
	Tolerance	$1e^{-2}$, $1e^{-3}$, $1e^{-4}$
	Solver	Auto, SVD, Cholesky, LSQR, Sparse CG, SAG, SAGA
Random Forest	Num. Estimators	50, 100, 200, 300, 400
	Criterion	MAE, MSE
	Max. Depth	16, 32, 64, 128, None
	Max. Features	Auto, sqrt, log2
SANN	Hidden Layer Shapes	(1,), (2,), (3,), (4,), (1,1,), (2,2,), (3,3,), (4,4,)
	Solver	LBGFS, Adam, SGD
	Alpha	0.00001, 0.0001, 0.001, 0.01, 0.1
	Initial Adam Learning Rate	1, 0.1, 0.01, 0.001
SVR	C	0.01, 0.1, 1.0, 10, 100
	Kernel	Linear, Polynomial, RBF, Sigmoid
	Degree	2, 3, 4, 5, 6
	Gamma	Scale, Auto
	Coef. 0	0.0, 0.5, 1.0
	Shrinking	True, False
	Tolerance	$1e^{-3}$, $1e^{-4}$, $1e^{-5}$
	Epsilon	0.0, 0.5, 1.0

Using the datasets described in Section II, the MAPE, calculated as per (2), is used to compare model performances. To facilitate greater transparency of model performance, Mean Absolute Error (MAE) scores are also included; see (3) – where y and $\hat{y}$ represent the actual and forecast variables respectfully.

$$MAPE = \frac{100\%}{n}\sum\left|\frac{y-\hat{y}}{y}\right| \quad (2)$$

$$MAE = \frac{1}{N}\sum|y-\hat{y}| \quad (3)$$

V. Results

Adopting the methodology detailed within Section III, results from the highlighted models are collated in Table 5 in order of MAPE score achieved. The models in Table 5 used a feature set of the 35 variables with the highest absolute Pearson Correlation scores. These models were used to perform a week-ahead forecast. Using the Lewis scale [19], detailed in Table 4, for MAPE evaluation, the top 9 scoring models all achieve a '*highly accurate forecast'*.

Table 4: Lewis scale for MAPE score evaluation

<10	Highly accurate forecasting
10-20	Good forecasting
20-50	Reasonable forecasting
>50	Inaccurate forecasting

Source: Lewis (1982, p.40) [15]

Table 5: Initial week-ahead forecast scores ordered by MAPE

Model	MAPE	MAE
Random Forest	4.72	628,853.78
Linear Regression	5.27	638,277.91
Lasso	5.73	651,323.98
Ridge Regression	6.36	648,035.18
KNR	6.43	818,972.62
Decision Tree	6.85	920,147.71
Bayesian Ridge	6.91	670,182.75
Extra Tree	7.38	880,740.06
Passive Aggressive	7.79	800,953.75
Linear SVR	12.32	1,633,582.38
SANN	13.40	1,498,485.48
NARX	58.48	13,181,710.14
Nu SVR	106.04	10,965,586.27
Gaussian Process	115.28	10,441,335.37
SVR	119.48	10,243,318.95

These error scores were then reduced further using feature scaling methods and other feature sets. The importance of experimenting with different scaling methods and feature sets are illustrated by Figures 3 and 4 respectively. In Figure 3, it can be seen that by scaling data appropriately, the forecasting models perform with lower error margins on average.

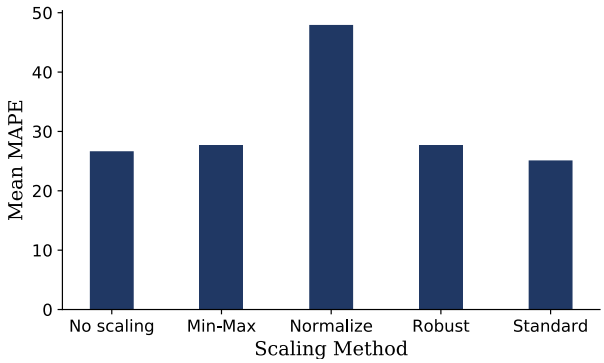

Figure 3: Mean MAPE by feature scaling method across all models

The five scaling methods used are: Min-Max scaling, no scaling, normalization, robust scaling, and standardization. These methods are implemented using scikit-learn[4]. The Min-Max scaler scales data into a fixed range, i.e. 0 to 1. Normalization transforms non-zero values using the l2 unit norm. The robust scaler is an alias for Interquartile Range (IQR) scaling. Data are scaled between the 1st and 3rd quartiles. This should theoretically be robust to outliers; hence the name. Finally, standardisation involves subtracting the mean and dividing by the standard deviation.

[4] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.preprocessing

978-1-6654-5227-4/22/$31.00 ©2022 IEEE

In Figure 4, one can observe that models produce better forecasts in using feature scoring methods that resulted in smaller feature sets, i.e. F-regression, Feature Importance and Mutual Information. The worst scores were observed when using all features.

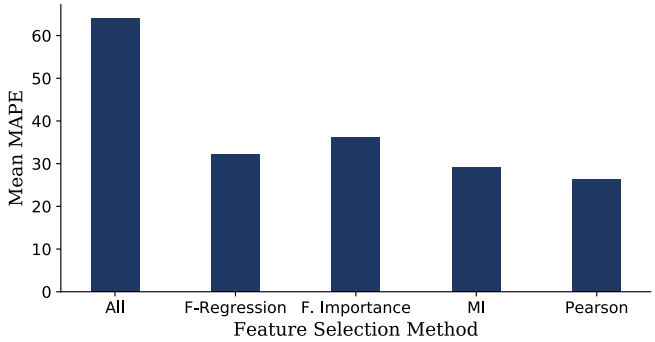

Figure 4: Mean MAPE by feature selection method for all models

The research undertaken in this study also experimented with resampling of data but did *not* find resampling to be beneficial to model performance. Under the assumption that there may be subtle pattern differences in post-quota data to peri-quota milk yields, data after the quota abolition was resampled. A binary label was applied to data to indicate whether a dataset record was post- or peri-quota. Observations were then resampled using the methods listed below:

- Adaptive Synthetic (ADASYN) sampling [20]
- No resampling
- Random over sampling
- Random under sampling
- SMOTE [21]
- SMOTE-ENN [22]
- SMOTE-Tomek [22]

These resampling methods were implemented using the Python imbalanced-learn[5] library. Presented in Figure 5, the results of using resampling on a subset of data indicated that resampling actually reduces model accuracy or has no effect.

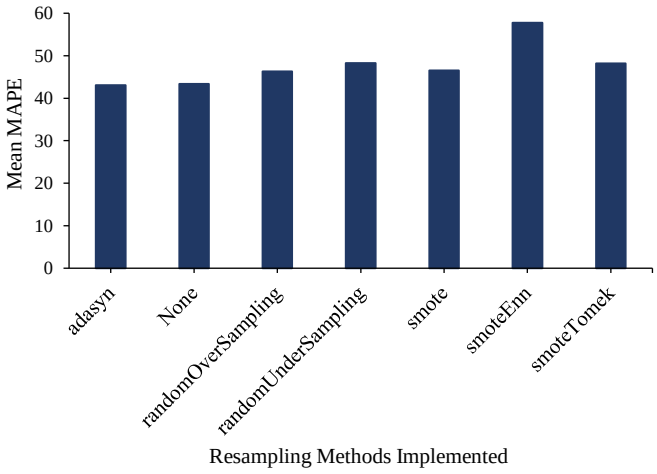

Figure 5: Mean MAPE by resampling method on a subset of data

Using feature scaling and alternative feature scoring methods, week-ahead forecasting scores were improved substantially. This fact is substantiated by the results presented in Table 6. Again, referring to the Lewis scale, the top 12 models all deliver '*highly accurate forecasts',* NARX achieves a *'good'* forecast, while the Nu SVR and SVR model can be categorized as *'inaccurate'*.

Table 6: Improved week-ahead forecasting scores

Model	Feature Scoring	Feature Scaling	MAPE
Random Forest	MI	Min-Max	4.63
MLR	F-Regression	Standard	5.27
Lasso	Pearson	Robust	5.73
KNR	Pearson	Min-Max	6.18
Gaussian Process	Pearson	Min-Max	6.33
Ridge Regression	Pearson	No scaling	6.36
Decision Tree	F-Regression	Standard	6.36
Linear SVR	MI	Min-Max	6.55
Extra Tree	F-Regression	Robust	6.95
SANN	Pearson	Standard	7.02
Bayesian Ridge	F-Regression	No scaling	7.37
Passive Aggressive	Pearson	No scaling	8.32
NARX	Pearson	Standard	19.21
Nu SVR	F-Regression	Robust	106.04
SVR	F-Regression	Robust	119.48

The outlined methodology can be extended to perform longer-term forecasts up to 52 weeks ahead. A selection of the models are inherently multi-output models, signifying that they may be configured to output more than one prediction. For example, a model with 10 outputs can be used to forecast 10 milk yields, representing that for the ensuing 10 weeks. For these models, along with non-multi-output models, two multi-model wrapper methods were implemented: multi-model and chain regression. The multioutput methods selected for each model are detailed in Table 7.

Table 7: Optimal multioutput method configurations

Model	Multioutput Method
Bayesian Ridge	Multi-model
Decision Tree	Single model
Extra Tree	Single model
Gaussian Process	Single model
KNR	Multi-model
Lasso	Single model
MLR	Multi-model
Linear SVR	Multi-model
NARX	Single model (Recursive)
Nu SVR	Chain-model
Passive Aggressive	Multi-model
Random Forest	Single model
Ridge Regression	Single model
SANN	Multi-model
SVR	Chain-model

[5] https://pypi.org/project/imbalanced-learn/

978-1-6654-5227-4/22/$31.00 ©2022 IEEE

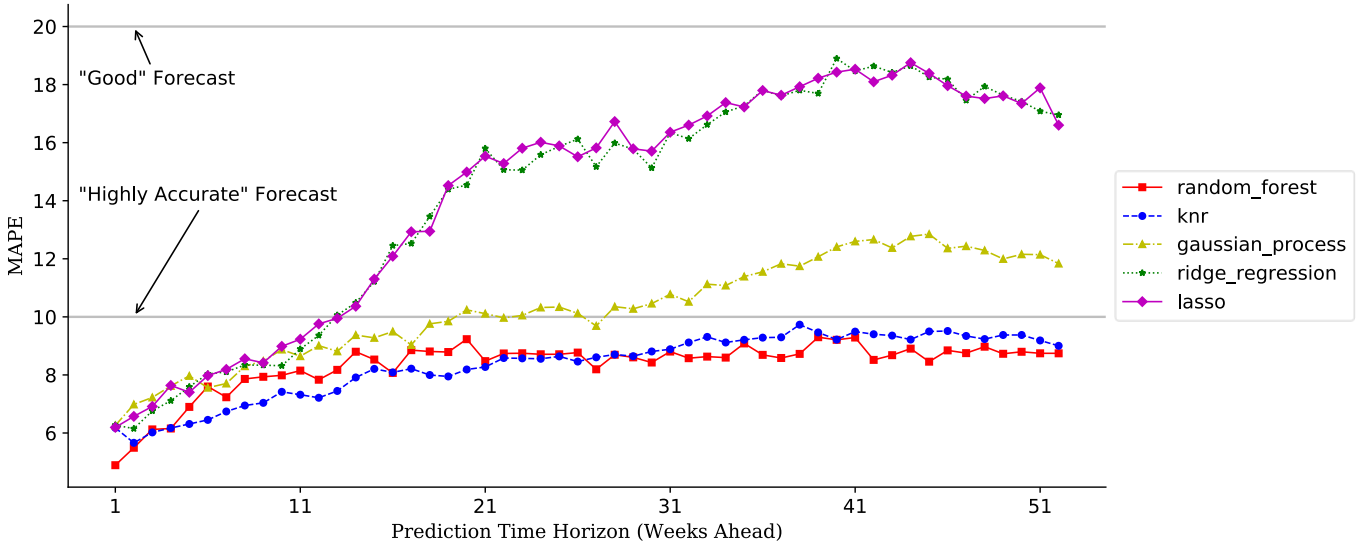

Figure 6: Mean MAPE for longer-term forecasts

In multi-model regression, a number of models are trained equal to the number of desired outputs, i.e. for a 52-week ahead forecast, 52 models are trained and produce outputs simultaneously. Chain regression operates in a similar manner. However, each model produces its outputs consecutively, where the output of each model in the 'chain' is used as an additional feature for the following model. The ML forecasting models were trained and tested using each of the available methods. The mean MAPE scores across all 52 weeks for all models are presented in Table 8.

Table 8: Mean MAPE averaged across increasing forecast horizons from 1 to 52 weeks for 30 iterations

Model	MAPE	MAE
Random Forest	8.28	1,138,221.46
KNR	8.35	1,077,417.30
Gaussian Process	10.27	1,165,771.13
Ridge Regression	14.16	1,610,928.18
Lasso	14.27	1,612,834.16
Extra Tree	14.47	1,964,233.40
Decision Tree	14.55	1,965,347.93
NARX	19.21	3,086,593.59
Linear SVR	21.58	2,098,389.72
SANN	22.57	2,564,965.40
MLR	27.29	2,551,649.15
Bayesian Ridge	37.47	4,025,336.03
Passive Aggressive	49.51	5,649,816.24
Nu SVR	111.73	11,022,258.18
SVR	125.98	10,264,124.99

Only the Random Forest and KNR models consistently perform '*highly accurate forecasts'* for the whole 52 weeks. In Figure 6, the mean MAPE scores for the top 5 performing models are plotted against for increasing prediction time horizon. While each of the 5 models perform '*highly accurate forecasts*' initially, the accuracy of the Random Forest and KNR models are visibly more reliable for longer term forecasts. Gaussian Process forecasts remain *'highly accurate'* until approximately week 19; most subsequent forecasts are classified as *'good'*. Ridge Regression and Lasso degrade to *'good'* at weeks 14 and 13 respectively. The works of Zhang et al. and Murphy et al. performed milk yield forecasts on Irish data for individual cows and herds [5, 8, 9]. They are the most similar studies with which we can compare model performance despite the difference in data granularity. To facilitate a more direct evaluation, the MLR, SANN and NARX models have also been implemented in this study. The MLR model is initially *'highly accurate'* for 11 weeks on average with a mean week ahead MAPE of 5.27 although its mean MAPE averaged across an expanding forecast horizon is 27.29 which can be classified as *'reasonable'*. The SANN model is more consistent overall with a *'reasonable'* mean MAPE of 22.57 while the NARX model is *'good'* at 19.21. This is consistent with the findings in [8] which claim that the NARX model is the most effective of the three overall, but neither model achieves the best scores in all situations [23]. Similarly, the Random Forest and KNR models alternate between best and second-best forecasting models depending on the forecast horizon. This is most noticeable for horizons in the range of 5 to 27 weeks ahead where KNR is more accurate on average. While the Random Forest is more accurate overall, the KNR model has a less noisy and flatter error curve. This increased consistency may stem from having a smaller hyperparameter search space and is therefore easier to optimize. The effectiveness of the Random Forest model here corroborates that found in [15].

VI. Conclusions

The purpose of this research is to examine the feasibility of using machine learning models for short- and long-term milk yield forecasting. Using the accepted Lewis scale for interpreting MAPE scores, it can be claimed that the Random Forest and KNR models execute '*highly accurate'* forecasts of milk yield up to one year in advance. The KNR model provides better forecasts for horizons in the range of 5 to 27 weeks, while the Random Forest model is slightly more accurate overall. Six other models are shown to initially produce *'highly accurate'* forecasts which degrade to *'good'* forecasts as the forecast horizon expands.

It was noted that feature scaling and feature selection improved the forecasts substantially. Conversely, resampling

978-1-6654-5227-4/22/$31.00 ©2022 IEEE

methods reduced the forecast accuracy. The inclusion of weather data was not found to be impactful, but previous milk yield values, numbers of suppliers, and some engineered features all displayed strong correlations. Future work could entail alternative options for modelling including deep learning, hybrid models and statistical approaches. The methodology used in this paper is generic enough that it should be applicable to other milk yield datasets with minimal alteration. The means of feature selection and model selection are data agnostic. For example, Koprinska et al. used the aforementioned feature scoring and ranking method to select features for forecasting electricity load [18]. The models we have used are implemented using scikit-learn which is a generic ML library. Thus, the methodology presented in this paper could even be applied to other forms of time series data.

VII. Acknowledgements

The authors would like to thank Kerry Group for its assistance in the study. In addition, they would like to express their gratitude to Enterprise Ireland for its support for a feasibility study in this domain. Finally, the authors wish to thank the Research Office within the Munster Technological University for its support.

VIII. References

[1] T. Lapple and D. Hennessy, "The capacity to expand milk production in Ireland following the removal of milk quotas," *Irish Journal of Agricultural and Food Research,* 2012.

[2] The Department of Agriculture, Food and the Marine, "Food Wise 2025," 16 12 2020. [Online]. Available: https://www.gov.ie/en/publication/a6b0d-food-wise-2025/. [Accessed 24 11 2021].

[3] S. Kavanagh, "Feeding the Dairy Cow," in *Teagasc Dairy Manual,* Teagasc, 2016.

[4] C. Moran, "Seasonality of Irish milk production becoming more pronounced," Agriland, 13 08 2016. [Online]. Available: https://www.agriland.ie/farming-news/seasonality-of-irish-milk-production-becoming-more-pernounced/.

[5] F. Zhang, P. Shine, J. Upton, L. Shaloo and M. D. Murphy, "A Review of Milk Production Forecasting Models: Past & Future Methods," ResearchGate, 2018.

[6] "Milk Statistics December 2020," Central Statistics Office, Ireland, 29 01 2021. [Online]. Available: https://www.cso.ie/en/releasesandpublications/er/ms/milkstatisticsdecember2020/.

[7] Holstein Foundation, "Milking and Lactation," November 2017. [Online]. Available: http://www.holsteinfoundation.org/pdf_doc/workbooks/Milking_Lactation_Workbook.pdf.

[8] M. D. Murphy, M. J. O'Mahony, L. Shalloo, P. French and J. Upton, "Comparison of modelling techniques for milk-production forecasting," *Journal of Dairy Science,* vol. 97, 2014.

[9] F. Zhang, J. Upton, L. Shalloo, P. Shine and M. D. Murphy, "Effect of introducing weather parameters on the accuracy of milk production forecast models," *Information Processing in Agriculture,* 2020.

[10] S. Soeharsono, S. Mulyati, S. Utama, W. Wurlina, P. Srianto and T. I. Restiadi, "Prediction of daily milk production from the linear body and udder," *Veterinary World,* vol. 13, 2020.

[11] R. S. Gandhi, D. Monalisa, V. B. Dongre, A. P. Ruhil, A. Singh and G. K. Sachdeva, "Prediction of first lactation 305-day milk yield based on weekly test day records using artificial neural networks in Sahiwal Cattle," in *Proceedings of the 5th Indian International Conference on Artificial Intelligence,* 2011.

[12] D. B. Jensen, M. v. d. Voort and H. Hogeveen, "Dynamic forecasting of individual cow milk yield in automatic milking systems," *Journal of Dairy Science,* vol. 101, no. 11, 2018.

[13] S. McParland, B. Coughlan, B. Enright, M. O'Keeffe, R. O'Connor and D. B. L. Feeney, "Prediction of 24-hour milk yield and composition in dairy cows from a single part-day yield and sample," *Irish Journal of Agricultural and Food Research,* 2019.

[14] A. Liseune, M. Salamone, D. V. d. Poel, B. v. Ranst and M. Hostens, "Predicting the milk yield curve of dairy cows in the subsequent lactation period using deep learning," *Computers and Electronics in Agriculture,* 2021.

[15] Q. T. Nguyena, R. Fouchereaub, E. Frénoda, C. Gerardc and V. Sinchollec, "Comparison of forecast models of production of dairy cows combining animal and diet parameters," *Computers and Electronics in Agriculture,* 2020.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher and M. Perrot, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* 10/11/2011.

[17] A. R. Gollou and N. Ghadimi, "A new feature selection and hybrid forecast engine for day-ahead price forecasting of electricity markets," *Journal of Intelligent and Fuzzy Systems,* 2017.

[18] I. Koprinska, M. M. Rana and V. G. Agelidis, "Correlation and instance based feature selection for electricity load forecasting," *Knowledge-Based Systems,* 2015.

[19] C. D. Lewis, A Radical Guide to Exponential Smoothing and Curve Fitting, Butterworth Scientific, 1982.

[20] H. He, Y. Bai, E. A. Garcia and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *IEEE International Joint Conference on Neural Networks,* 2008.

[21] N. V. Chawla, K. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research,* 2002.

[22] R. C. P. M. C. M. Gustavo E. A. P. A. Batista, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter,* 2004.

[23] F. Zhang, M. D. Murphy, L. Shalloo, E. Ruelle and J. Upton, "An automatic model configuration and optimization system for milk production forecasting," *Computers and Electronics in Agriculture,* 2016.

978-1-6654-5227-4/22/$31.00 ©2022 IEEE