Error Estimation for Output Prediction of Photovoltaic Systems

REINIER L. C. VAN DER HORST, University of Twente, The Netherlands

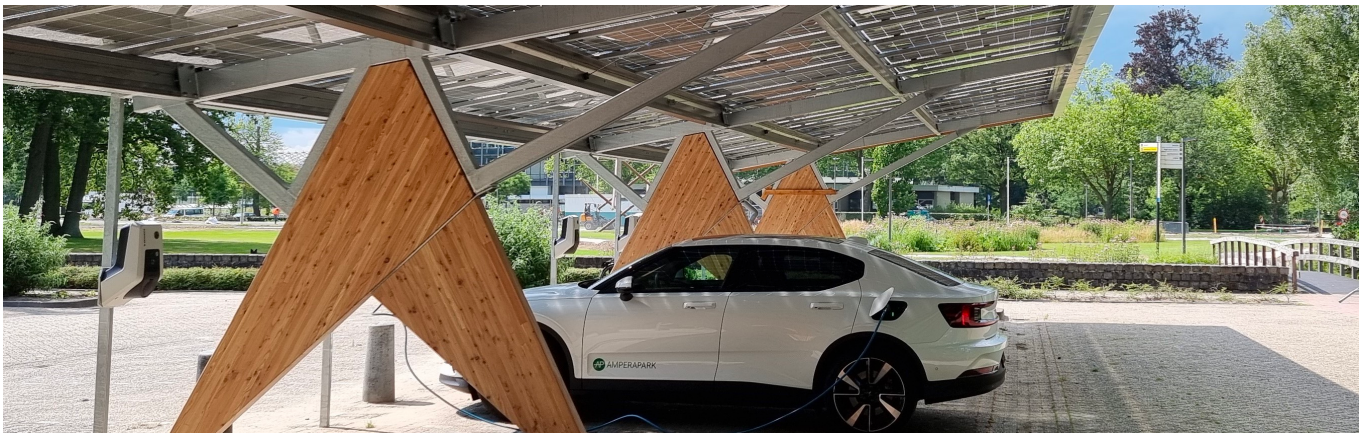


Fig. 1. SlimPark test location on UTwente[16]

Predictions for the power output of renewable energy sources are not always accurate. Gaining insight in the error of predictions can help grid operators manage the power grid more efficiently. This is especially important now that common households produce their own energy though Photovoltaic (PV) systems more frequently. Currently, the increase in the amount of energy generated through PV systems already leads to congestion and damage to the main energy grid. To counter this development, more anticipating control is required, which in turn requires more insight into future energy generation. In this research paper, the viability of an independent centralised model that estimates the error in the power output predictions made by a PV system is analysed. Multiple Linear Regression and XGBoost are trained on weather data in order to estimate the error of a PV prediction model. Machine Learning models prove to be a viable tool to provide insight into the reliability of output predictions, especially in classifying probable over-/under-estimations.

Additional Key Words and Phrases: Photovoltaic, Output Prediction, Error Estimation, Solar Energy, PV

1 INTRODUCTION

For the last few years, the energy system has shifted more towards sustainable and renewable energy sources. One of the promising technologies in this transition is the integration of solar power through Photovoltaic (PV) cells, more commonly known as solar cells [25]. The reduction in fabrication costs and increase in efficiency and durability have made PV systems a popular renewable alternative for fossil fuel based energy generation [9].

The intermittent nature of the power production of PV systems prevents these systems from being used as the main mode of generating energy for our current electrical grid. Intermittent power production can damage our current energy grid by creating voltage and frequency anomalies or excessive loads during peak production. Besides this, there is the fact that traditional power plants can supply a consistent and predictable power output that can be stepped up or throttled down to fit the specific energy demands at any moment. While solar energy can be throttled just like traditional energy sources, the maximum output cannot be stepped up due to dependency on uncontrollable factors like the local solar irradiance values. These fluctuations have become more common now that average households also have access to these technologies instead of relying on a few big energy generation sites spread across the country. This results in a more decentralised energy market with bi-directional flow which consequently leads to hard to predict energy output from a plethora of suppliers.

Fluctuations in power output in our current electrical grids give rise to an imbalance in supply and demand which forces grid operators to find solutions to counter the effects of this imbalance [20, 21, 23, 24]. The increase in the number of suppliers that mostly supply an inconsistent amount of energy to the grid, increases the complexity of this problem. Furthermore, these suppliers might predict their outputs using different methods and metrics with varying rates of success. This consequently complicates managing the supply and demand of the main energy grid even further. Having a robust energy management system can aid in reducing imbalances and damages caused by these fluctuations by, for example, utilizing energy storage solutions or the flexibility inherent to demand.

Because the accuracy of the models used to predict PV output can differ between methods and implementations of energy management systems, it is hard to allocate resources to guarantee grid

TScIT 37, July 8, 2022, Enschede, The Netherlands

stability effectively. Besides that, the data and models on which the predictions are based and the performance metrics thereof are generally not easily accessible if they are available at all. In order to manage the main energy grid it would be useful to be able to estimate the inaccuracy of a given model in certain weather conditions, as this indicates to what extend backup resources need to be available. This situation gives rise to the following research question:

Can the error of a PV output prediction model be estimated using a different and independent model with different data sources than the prediction model?

To test whether the error of such an output prediction model can be estimated, machine learning will be utilised to find relations between weather data and the error in the output prediction of the original model. In this report the performance requirements, the methodology of creating this model, and results are outlined. Subsequently, the results are analysed in order to answer whether this approach is viable when trying to predict estimation error.

2 REQUIREMENTS

In order for this approach to be considered viable it must meet the following three requirements:

- **Accurate:** The error estimation must be accurate enough to improve the final output estimation if the error is taken into account in post processing of the prediction.
- **Independent:** The approach must be generally applicable to models that use unknown data and approaches to realize their predictions.
- **Flexible:** Since weather, and thus PV systems, perform dependent on location specific conditions it is important that this approach can make use of the possibly limited data available.

The requirements that a real life implementation would have to meet are not noted in this list because functional real life implementations are outside the scope of this research.

3 RELATED WORK

Research on how estimations for PV output can be obtained is important when we look at estimating the error of these models. After all, only measurements related to PV output are to be considered since unrelated data might skew the estimation.

Most estimation models that were published by studies conducted nearly ten years ago use stochastic methods in order to estimate the irradiance level in the future. Predictions are made on the basis of camera images that predict predict the stochastic movement of clouds in front of the sun [5], stochastic state spaces [17] or irradiance values of other nearby PV systems [11].

In more recent years, artificial intelligence based prediction methods have become a popular research topic. Estimation methods based on decision trees [15], neural networks [3, 12] and deep learning [4] can be found, which generally yield better results than the older methods.

There is very little research on predicting the error of models. One possible explanation of this is the fact that in general, being able to predict the error of a model using certain features means that you can improve the accuracy of the original model by adding these features. In the case of the embedded models used to estimate output of PV systems, however, it might be hard or undesirable to update these models.

A common way to represent the accuracy of machine learning models are the Mean Absolute Error (MAE), Mean Square Error (MSE) and Root Mean Square Error (RMSE) scores. This can be a problem as these scores generalize the performance to a single value independent of external conditions. While this is a good way to estimate the general accuracy of a model, it does not provide enough information to aid in the task of managing energy since this metric cannot be used to compensate mistakes in individual measurements.

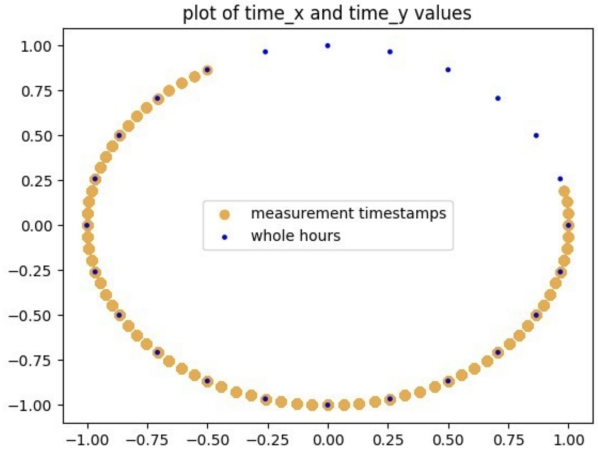


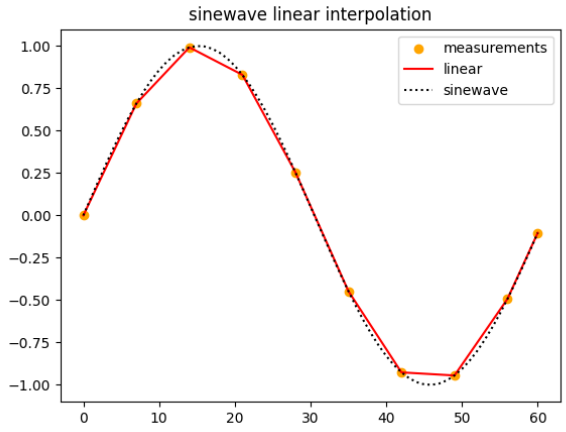
Fig. 2. Scatter plot of all time coordinates

4 METHODOLOGY

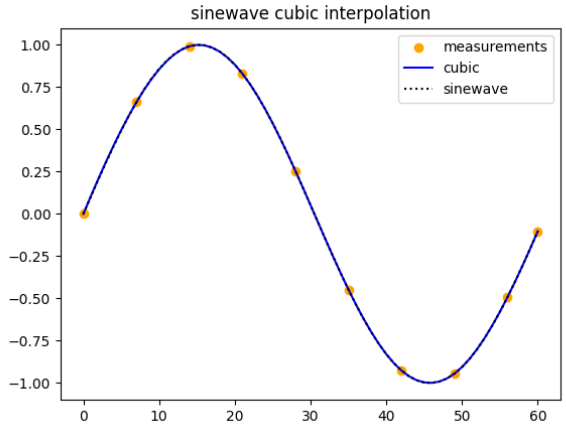
As mentioned before, machine learning models will be utilised to find a relation between weather data and the error in the output prediction of a PV system. The energy research group of the University of Twente manages a living lab facility called SlimPark, which is equipped with a PV system, electric vehicle chargers and a battery. This test location uses machine learning in the form of Multiple Linear Regression in order to estimate the PV output [8]. This implementation and the obtained historical data is used as a source of error data for the analysis.

4.1 Data acquisition

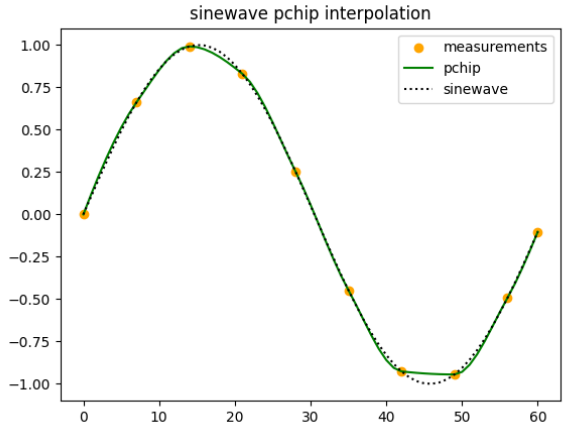
Predictions and actual output values recorded from February through June by the SlimPark PV system were taken as a source of error data. This PV system predicts its output based on irradiance forecast predictions. In order to get the weather data required to find possible relations between weather conditions and the error of the system the data published by the Royal Netherlands Meteorological



(a) Sinewave approximated with linear interpolation



(b) Sinewave approximated with cubic spline interpolation



(c) Sinewave approximated with PCHIP interpolation

Fig. 3. sinewave interpolations

Institute (KNMI) was taken. This is historical weather data instead of historical weather forecast data since historical weather forecast data for the location of the SlimPark PV system was not readily available. The SlimPark installation does not have a weather station on site. In order to keep the weather data as consistent as possible with the weather at the SlimPark site, the weather station called "twenthe airport" was used as a source of weather data, because this was the closest weather station that the KNMI monitors.

4.2 Data Preparation

The data gathered from the KNMI weather station is recorded every 10 minutes whereas the historical data of the SlimPark PV system is recorded every 15 minutes. This causes half of the data obtained through SlimPark to be unsynchronizable without prepossessing. The data point on the 15^{th} and 45^{th} minute of each hour can not be matched directly to a data point taken from the KNMI data set. In order to keep our data set as big as possible, the data from the KNMI data was interpolated in order to obtain feasible data points for the 15^{th} and 45^{th} minute of every hour.

4.2.1 Interpolation.

Linear interpolation [19] was considered, but due to the inability to represent the smooth transitions of some of the data points in the weather data effectively this method was not selected (see figure 3a), even though a form of linear regression is used to calculate average monthly temperatures according to a literature review commissioned by the KNMI [22]. A cubic spline [14] was considered next since this method generates smoother curves than linear interpolation, as can be seen in figure 3b. This method was not selected due to the tendency of some of the data points in the weather data to plateau. The Gibbs phenomenon would arise using splining [18], which is undesirable. Apart from that, the values generated by cubic splining can be higher or lower than the measured values it lies between. This can be a problem when interpolating, for example, rain gauge measurements. If it stops raining suddenly the cubic splines interpolated values can give a measurement of negative rainfall, which is impossible. Lastly, different splining methods are considered a viable method of interpolating weather data [10, 13], but this is only the case when considering data with resolutions of a day or lower according to the same KNMI literature review [22]. The interpolation method that was selected after these considerations was a "Piecewise Cubic Hermite Interpolating Polynomial" (PCHIP) [7]. It works by calculating a cubic polynomial that goes through the two neighbouring points at a specific slope in order to smoothly transition point to point. This method is smoother than linear interpolation, but always remains in the range of the given measurements that it is surrounded by, as can be seen in figure 3c.

4.2.2 Data Selection.

The data taken from the KNMI publications contains data points of different time resolutions. An example of this would be rainfall in the last 24, 12, 6 and 1 hour(s). The data points with a time resolution closest to 15 minutes has been taken in order to keep the resolution of the model and the other data points as similar as possible. In the case of the above example the data point for rainfall in the last hour would be taken. This target resolution was chosen since the data

resolution of the SlimPark data set is 15 minutes.

At some points the sensors of the weather station were partially turned off. These entries of the KNMI data have been removed, as well as the interpolated values around these data points. There are also extreme outliers, which were also removed. Extreme outliers, in this context, are defined as values that are more than 5 standard deviations away from the mean, These outliers were removed because they coincided with moments of sensor outage, indicating that the measurements might be unreliable.

After this data selection, the final data set contains data from the 12^{th} of February 2022 up to and including the 24^{th} of June 2022 with no usable data on a few days at the end of march and the middle of June. The set consist of 6912 data points, containing 119 features in total. The weather data consists of features such as rainfall in the last hour, global solar irradiation, wind speeds, maximum and minimum temperatures, weather condition codes, air pressure, and the prediction and output data of the SlimPark PV system.

4.2.3 *General Data Manipulation.*

Since the SlimPark test location only provides prediction values and the actual output, the prediction error needs to be calculated from these values. Prediction and output values provided by the living lab facility are always negative because energy generation is defined as negative energy consumption. Because of that, we use the following formula to calculate the prediction error:

```
Prediction Error = Real Output - Predicted Output
```

Calculating the prediction error in this way results in negative numbers when the prediction model underestimates and positive values when the model overestimates. The unit Watt_τ is defined to be equal to 1 Watt/15m in order to be able to represent the. This means that 4 Watt_τ is equal to 1 Wh exactly.

When taking into account general data manipulations, the following changes have been made: there is no energy generation during the nights, so the weather and SlimPark data have been removed before sunrise and after sunset. The time of day is linked to the angle at which sun rays hit the PV panel surface, which makes time an important variable as well. The timestamps of the SlimPark and KNMI data sets both use the 24 hours time notation. In order to make this a normalized value, the time was plotted on a circle (see figure 2) using the following algorithm:

```
time_numeric = hours*100 + (minutes/60)*100
timestamp_x = sin(2.0 * pi * time_numeric / 2400)
timestamp_y = cos(2.0 * pi * time_numeric / 2400)
```

Here timestamp_x and timestamp_y represent the x and y coordinates of the timestamp in a graph. This approach was chosen so that differences in time of two time stamps are proportional to the distance of the coordinates of these same two timestamps.

For the categorical values, a one-hot-encoding was used in order to be able to use this information in the models as well. This means that every category that is present in the data will be a separate feature with an indication of 1 if the data point is categorised in that specific category and 0 otherwise.

4.3 Models

In order to gain more insight in the usefulness of this approach we aim to find relations between the weather data and the error of the predictions made by the PV system prediction model. To be exact, weather data will be fed into machine learning models in order to estimate the error of an output prediction using the following machine learning algorithms:

- **Multiple Linear Regression (MLR):** Since this method is also used in the PV prediction algorithm [8], this algorithm has proven itself capable in a very similar situation. This makes this algorithm a good starting point. The 'scikit-learn' (sklearn) Python library [1] was used in order to provide en efficient implementation of Multiple Linear Regression.
- **XGBoost:** XGBoost [6] is a tree boosting method that has proven quite capable in machine learning competitions. Because of good performance in competitions XGBoost is expected to yield good results. The 'xgboost' Python library [2] was used in order to provide an efficient implementation of this algorithm.

4.4 Evaluation

The implemented machine learning algorithms are evaluated using the following metrics:

- Mean Absolute Error (MAE)
- Mean Square Error (MSE)
- Root Mean Square Error (RMSE)
- Percentage of correct over-/under-estimation indication (PS)
- Percentage of better predictions after compensation (PB)

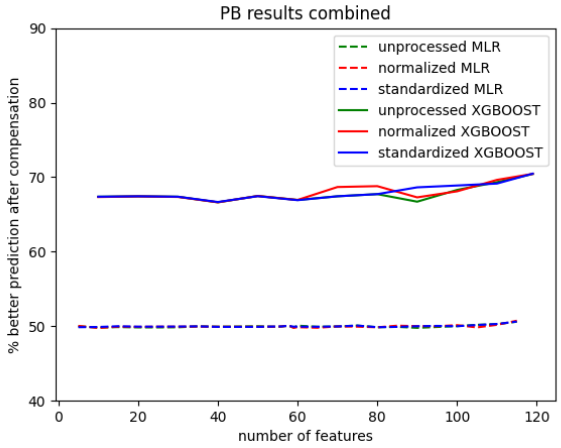
MAE is chosen since it gives a intuitive indication of the magnitude of the error of a model. MSE and RMSE are common indications of the performance of models and the MSE tends to enlarge difference between errors, making it useful in order to see small differences in inaccuracies more clearly. PS and PB are added in order to gauge the potency and reliability of the method.

Three data formats were used in order to evaluate the models. Firstly, we take the data from the selected data set without any further preprocessing beyond the data preparation and selection. This data set is called the unprocessed data set. Secondly, we have a standardised data set. This approach was taken in order to limit the influence of having various units and scales within the data set. Coincidentally, this should theoretically improve prediction accuracy if the machine learning algorithm assumes a Gaussian distribution. Lastly, we have the normalised data set. This data set normalises all values so they have a value between 0 and 1.

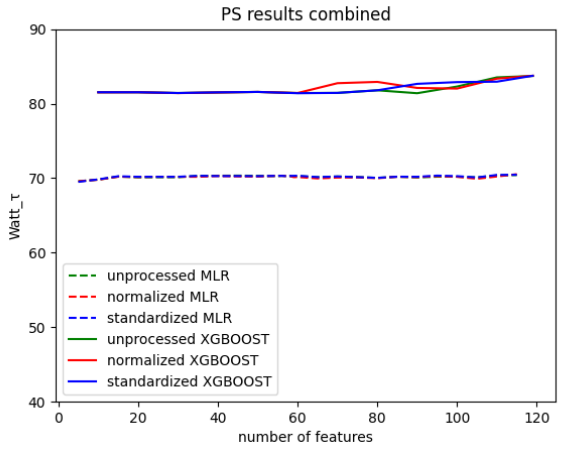
Some features might be more important than others. In order to see whether or not this is the case, the model is trained on different sets of features. These feature sets are calculated using a sequential feature selector, which is also provided by the scikit-learn library [1]. The sequential feature selector is configured to use forward selection, greedily choosing the feature with the most impact on

the desired outcome at each step.

The MLR algorithm with steps of 5 features at a time while the XGBoost algorithm was evaluated with 10 steps each time. This approach is chosen because of time constraints and the fact that a XGBoost model takes significantly longer to train.



(a) PB of MLR versus XGBOOST



(b) PS of MLR versus XGBOOST

Fig. 4. PB and PS metrics

To keep the evaluations consistent and reliable, a 10-fold cross validation is used. This means that the data set is cut up in 10 parts, of which 9 parts are taken in order to train the models and the 10^{th} part is used in order test the accuracy of the model according to the aforementioned metrics. Consequently this is done 9 more times. After this, the average of these 10 results is taken as an indication of the performance of that configuration.

5 RESULTS

The MAE, MSE, and RMSE of the prediction model, MLR, and XGBoost are plotted in figure 5. The output prediction model is untouched in this research, however, some of the same performance metrics can be calculated for the results of this model. Including these metrics in these figures serves the purpose of providing a sense of scale to the metrics derived from the other models. This metric can be used in this way since calculating the difference between the error estimation and the actual error results in compensated prediction error (CPE). The compensated prediction error is defined as follows:

```
Prediction + Estimated Error = CPE
```

The PB and PS metrics cannot be taken from the original data, because these values are based on the relation between the error estimation and the output prediction, making PB of the prediction model always zero and the PS undefined. The PS and PB scores of the MLR and XGBoost algorithm can be seen in figure 4.

5.1 Multiple Linear Regression

Tables 1, 3, and 5 in appendix B contain the numerical values of the metrics of the Multiple Linear Regression model resulting from the different number of features obtained through sequential feature selection. In appendix A.1 the values of all of the different metrics for the Multiple Linear Regression are plotted in more detail in order to see the small differences resulting from the different pre-processing methods more clearly.

As can be observed in the above-mentioned plots an tables, the Multiple Linear Regression algorithm does not show significant difference between the performance between the unprocessed, standardised, and normalised set. It can clearly be seen that the algorithm performs the best when close to all features are included, even though the accuracy decreases slightly when more features are added up to 110 features. There is also a notable spike with an increased error on 75 features.

Since there is only little difference between the different number of features, additional measurements are taken between 55 and 60 features. In this interval the model is analysed upon adding each feature to see if there was more fluctuation in the inaccuracy when looking at a higher resolution. It can be seen that the accuracy tends to decrease slowly but steadily and there is no significant difference in the magnitude of the fluctuations.

Note that the number of predictions that have an increased accuracy after compensating for the estimated error is roughly 50% as can be seen in figure 4a. This fact combined with the fact that the MAE decreases by roughly 15% (see figure 5a) in relation to the MAE of the output prediction model, indicates that the accuracy improvement of predictions with a better CPE is larger than the accuracy reduction in the predictions with a worsened CPE. The RMSE and MSE metrics penalize errors further from the true value significantly harder than errors close to the mean, since these metrics are based upon the square of the error. Consequently, we can conclude that if the RMSE decreases more percentage-wise than the MAE, the peaks in the

error are either lower or significantly less frequent Besides this, it can be observed in figure 4b that the algorithm classifies an over- or under-estimation correctly roughly 70% of the time.
This in combination with the PB score of roughly 50%, means that the error estimation tends to overestimate the magnitude of the error.

5.2 XGBoost

Tables 2, 4, and 6 contain the metrics of the XGBoost model with several different number of features obtained through sequential feature selection. In appendix A.2 the values of all the different metrics for the XGBoost are plotted to more easily see the small differences between the different data formats.

When analysing the difference between the unprocessed, standardised, and normalised sets in figure 6, it can be observed that there is difference in performance. Even though there a difference in performance between the data sets at different points there is no data set that performs the best in all or nearly all cases.

It can be seen that the XGBoost algorithm performs better when more features are added. This is especially noticeable in the metrics derived from training with more than 60 features. Similarly to the results of the Multiple Linear Regression Algorithm, the XGBoost has a significant boost in performance when using all the features available to it.

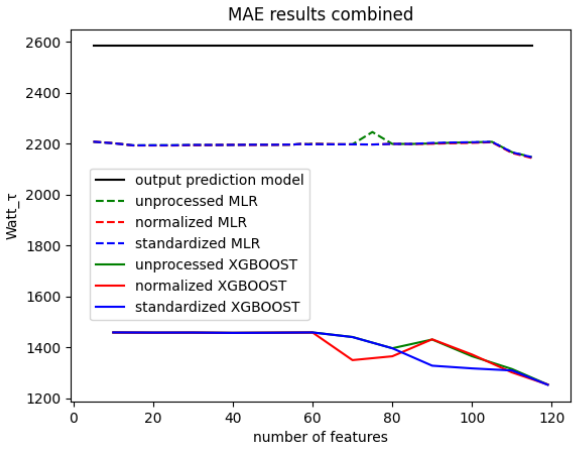
XGBoost clearly has better results than the multiple linear regression. With both more than 70% of predictions being improved when taking this error estimation into account, and over- or under-estimation classification accuracy comfortably above 80%, it can be seen that XGBoost is a potent tool when trying to estimate the error of PV output predictions. The MAE and RMSE have very similar accuracy increases and decreases, signifying that there are less fluctuations in prediction error.

6 DISCUSSION

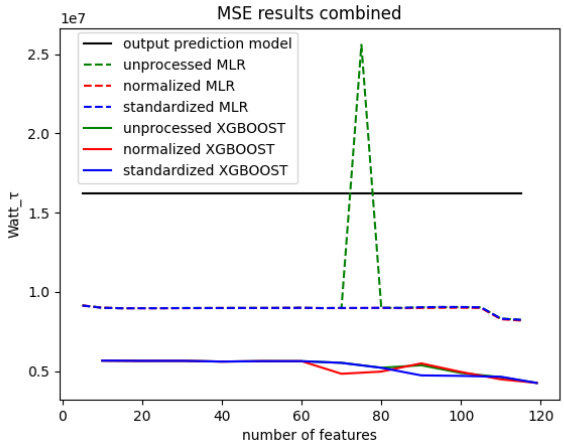
6.1 Integrity of Results

Even though the research is done in a controlled environment as much as possible, there are factors that might have impacted the integrity of the results.

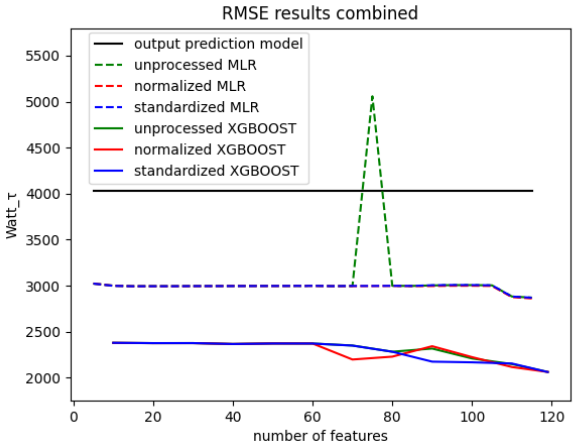
Firstly, evaluating the results of the models some extreme outliers were detected when looking at the inaccuracy of the MLR algorithm. The noticeable outliers had a MAE of more than 12000 Watt_τ and MSE of in the ranges of 0.5-2.0*10^{21} Watt_τ. After investigating the cause of these anomalies, it was discovered that these outliers were caused by a state change in the tooling used to keep the computer used to calculate the results awake during calculations. While these outliers are recalculated and now the presented results are more in line with expectations, it is unclear how many of the validation sets are impacted in a less severe manner. The error is completely reproducible and an alternative method of keeping the computer awake has been considered. But due to time constraints, recalculating the entire data set is impossible.



(a) MAE of all methods



(b) MSE of all methods



(c) RMSE of all methods

Fig. 5. MAE, MSE, and RMSE metrics

It is also unclear in what capacity the XGBoost results are impacted by this issue, since these results were also generated while using the same tooling.

Secondly, the data that was used in order to get these results is not perfectly representative of a real life situation. The data that was used is historical weather data instead of historical weather forecast data. It is thus possible that the results from a practical application of this method will have different, most likely worse, results. The results could also be skewed because the weather station is not at the same location as the PV system. Half of the data is not an actual measurement but rather interpolated, which might cause these data to not be representative of the actual weather conditions at that time.

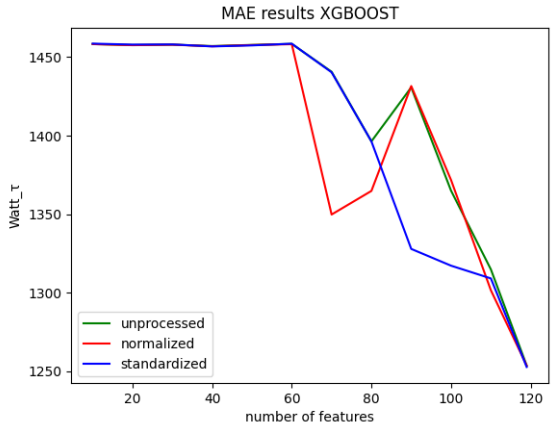


Fig. 6. MAE scores of the XGBoost model for different data sets

6.2 Multiple Linear Regression

When looking at the results for the MLR algorithm, there is very little difference between the unprocessed, standardised and normalised results. This is quite remarkable since linear regression assumes a Gaussian distribution, so it can be expected that using standardised or normalised data sets would result in better predictions than using the unprocessed data set. Since this is not the case it is hypothesised that the library from which the linear regression algorithm is taken detects when the data is not normalised or standardised and pre-processes the data accordingly. Researching this falls outside the scope of this research.

The fact that the MLR error estimation does not yield significantly better results can possibly be attributed to the fact that the there might not be a linear relation between the weather data and the prediction error. It can, however, be seen that there is useful information contained in the weather data points. The CPE has a lower Mean Absolute error than the original prediction and an over-/under-estimation can be classified with a higher accuracy than 50%. The outlier in the error of the unprocessed error at 75 features can likely be attributed to the processing errors caused by the tooling mentioned in paragraph 6.1 and thus will not be considered as a representative data point.

6.3 XGBoost

The results of the XGBoost model seem to be as expected. It can be seen that the normalised and/or standardised data set yield similar or better results than the unprocessed data set in most cases. This can possibly be explained by the fact that XGBoost utilises linear boosting among other things. Linear boosting assumes a Gaussian distribution, just like linear regression, which consequently solidifies the hypothesis that the Multiple Linear Regression model that is used likely pre-processes data that is not normalised or standardised.

The performance of the XGBoost model can quite likely be improved further by fine-tuning the algorithm. This can be done by tweaking the number of estimators, learning rate, max depth, and the number of scolumns considered for building trees. There is also a XGBoost classifier instead of a regressor, which could possibly improve the performance on over-/under-estimation classification.

7 CONCLUSION

While the MLR model only improves the output prediction roughly 50% of the time, the magnitude of the average error does decrease significantly with 15%. When looking at the XGBoost model, the error reduction gets even better with a magnitude decrease of almost 50% . This method is also more reliable with around 70% of predictions having a lower CPE than the original prediction error. The classification is both potent using Multiple Linear Regression and XGBoost with an accuracy of roughly 70% and 82% respectively. Both models seem to perform when using different feature sets signifying that there is a wide range of weather data that can be used using this approach. The initial data used to derive the output prediction was not used to estimate the error, thus this method is only reliant on weather data, the output prediction of a given model, and the actual output of the PV system when training. Only relying on weather data and the output prediction, when estimating the error.

Consequently we can say that this approach is accurate, flexible and independent which gives us enough information to answer the research question: Can the error of a PV output prediction model be estimated using a different and independent model with different data sources than the prediction model?

Yes, the error of a PV output prediction model be estimated using a different and independent model with different data sources than the prediction model. The approach proposed in this research is potent, but further research needs to be done in order to explore the extent of its potential and its limits.

8 FUTURE WORK

Follow-up research topics that could provide more insight into the potency, reliability and limits of this approach could be:
Firstly, there could be other feature selection methods considered since the current feature selection method that was used did not

seem to have much effect. This means that there could be significant improvements if an effective feature selection method were to be implemented. There are feature combinations that perform better than others as can be seen at in at the 110 feature data points. The greedy sequential feature selection might not be the most suitable method to find these combinations.

Secondly, other machine learning models could be used in order to see if there are other models that perform better than the ones that were tested in this research. Separating classification of over-/under-estimation from the estimation of the magnitude could be researched to see whether this yields better results than having one combined estimation.

Thirdly, it could be researched whether placing a weather station closer to the PV-system causes this approach to yield better results as well as seeing the impact of using forecast data instead of historical weather data to train the models. In the same sense, weather forecast data can be used in order to research the potency of the approach in practical situations.

Finally, because the data in the data set is collected over a period of 5 months, the effect of changing seasons could not be evaluated. Further research with data collected over multiple years could thus be prove to be valuable.

ACKNOWLEDGMENTS

I am grateful for my supervisors who have helped me with cleaning my data and steering my research in the right direction. Besides that, I want to give special thanks to M. F. Verkleij who helped me by providing me with weather data in order to create the data set I have used during this research.

REFERENCES

[1] 2022. SciKit Learn Python library. https://scikit-learn.org/stable/index.html. version = 1.1.1.

[2] 2022. XGBoost Python library. https://xgboost.readthedocs.io/en/stable/. version = 1.6.1.

[3] F. Alassery, A. Alzahrani, A. Khan, K. Irshad, and S. R. Kshirsagar. 2022. An artificial intelligence-based solar radiation prophesy model for green energy utilization in energy management system. *Sustainable Energy Technologies and Assessments* 52 (2022).

[4] O. Bamisile, A. Oluwasanmi, S. Obiora, E. Osei-Mensah, G. Asoronye, and Q. Huang. 2020. Application of deep learning for solar irradiance and solar photovoltaic multi-parameter forecast. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects* 0, 0 (2020), 1–21. https://doi.org/10.1080/15567036.2020.1801903 arXiv:https://doi.org/10.1080/15567036.2020.1801903

[5] D. Cai, T. Xie, Q. Huang, and J. Li. 2014. Short-term solar photovoltaic irradiation predicting using a nonlinear prediction method. In *IEEE Power and Energy Society General Meeting*, Vol. 2014-October.

[6] T. Chen and C. Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[7] F. N. Fritsch and R. E. Carlson. 1980. Monotone Piecewise Cubic Interpolation. *SIAM J. Numer. Anal.* 17, 2 (1980), 238–246. http://www.jstor.org/stable/2156610

[8] M. E. T. Gerards and J. L. Hurink. 2019. PV Predictions Made Easy: Flexibility Through Simplicity. In *Proceedings of the 25th International Conference on Electricity Distribution (CIRED 2019) (CIRED Conference Proceedings)*. CIRED. http://www.cired2019.org 25th International Conference and Exhibition on Electricity Distribution, CIRED 2019, CIRED ; Conference date: 03-06-2019 Through 06-06-2019.

[9] M. Gul, Y. Kotak, and T. Muneer. 2016. Review on recent trend of solar photovoltaic technology. *Energy Exploration & Exploitation* 34, 4 (2016), 485–526. https://doi.org/10.1177/0144598716650552 arXiv:https://doi.org/10.1177/0144598716650552

[10] C. H. Jarvis and N. Stuart. 2001. A comparison among strategies for interpolating maximum and minimum daily air temperatures. Part II: The interaction between number of guiding variables and the type of interpolation method. *Journal of Applied Meteorology* 40, 6 (2001), 1075–1084.

[11] I. Jayawardene and G. K. Venayagamoorthy. 2016. Spatial Predictions of Solar Irradiance for Photovoltaic Plants. In *2016 IEEE 43RD PHOTOVOLTAIC SPECIALISTS CONFERENCE (PVSC) (IEEE Photovoltaic Specialists Conference)*. IEEE, 267–272. 43rd IEEE Photovoltaic Specialists Conference (PVSC), Portland, OR, JUN 05-10, 2016.

[12] J. Li, H. Niu, F. Meng, and R. Li. 2022. Prediction of Short-Term Photovoltaic Power Via Self-Attention-Based Deep Learning Approach. *Journal of Energy Resources Technology, Transactions of the ASME* 144, 10 (2022).

[13] Z. Luo, G. Wahba, and D. R. Johnson. 1998. Spatial–temporal analysis of temperature using smoothing spline ANOVA. *Journal of Climate* 11, 1 (1998), 18–28.

[14] M. Marsden. 1974. Cubic spline interpolation of continuous functions. *Journal of Approximation Theory* 10, 2 (1974), 103–111. https://doi.org/10.1016/0021-9045(74)90109-9

[15] J. Moon, Z. Shin, S. Rho, and E. Hwang. 2021. A Comparative Analysis of Tree-Based Models for Day-Ahead Solar Irradiance Forecasting. In *2021 International Conference on Platform Technology and Service, PlatCon 2021 - Proceedings*.

[16] Univerity of Twente. 2021. UT campus AmperaPort. https://twente.energy/assets/img/utcampus-amperaport.jpg [Online; accessed July 3, 2022].

[17] M. Olama, A. Melin, J. Dong, S. Djouadi, and Y. Zhang. 2017. Stochastic short-term high-resolution prediction of solar irradiance and photovoltaic power output. In *2017 North American Power Symposium, NAPS 2017*.

[18] F. B. Richards. 1991. A Gibbs phenomenon for spline functions. *Journal of Approximation Theory* 66, 3 (1991), 334–351. https://doi.org/10.1016/0021-9045(91)90034-8

[19] M. K. Samarin. 2012. Linear interpolation. https://encyclopediaofmath.org/wiki/Linear_interpolation

[20] S. S. Sami, M. Cheng, J. Wu, and N. Jenkins. 2018. A virtual energy storage system for voltage control of distribution networks. *CSEE Journal of Power and Energy Systems* 4, 2 (2018), 146–154. https://doi.org/10.17775/CSEEJPES.2016.01330

[21] Q. Shi, H. Cui, F. Li, Y. Liu, W. Ju, and Y. Sun. 2017. A hybrid dynamic demand control strategy for power system frequency regulation. *CSEE Journal of Power and Energy Systems* 3, 2 (2017), 176–185. https://doi.org/10.17775/CSEEJPES.2017.0022

[22] R. Sluiter. 2009. Interpolation methods for climate data: literature review. *KNMI intern rapport, Royal Netherlands Meteorological Institute, De Bilt* (2009).

[23] Y. Sun, Z. Zhao, M. Yang, D. Jia, W. Pei, and B. Xu. 2020. Overview of energy storage in renewable energy power fluctuation mitigation. *CSEE Journal of Power and Energy Systems* 6, 1 (2020), 160–173. https://doi.org/10.17775/CSEEJPES.2019.01950

[24] Q. Tabart, I. Vechiu, A. Etxeberria, and S. Bacha. 2018. Hybrid Energy Storage System Microgrids Integration for Power Quality Improvement Using Four-Leg Three-Level NPC Inverter and Second-Order Sliding Mode Control. *IEEE Transactions on Industrial Electronics* 65, 1 (2018), 424–435. https://doi.org/10.1109/TIE.2017.2723863

[25] V. V. Tyagi, N. A. A. Rahim, N. A. Rahim, A. Jeyraj, and L. Selvaraj. 2013. Progress in solar PV technology: Research and achievement. *Renewable and Sustainable Energy Reviews* 20 (4 2013), 443–461. https://doi.org/10.1016/J.RSER.2012.09.028

A APPENDIX: DETAILED METRICS PLOTS

A.1 Multiple Linear Regression

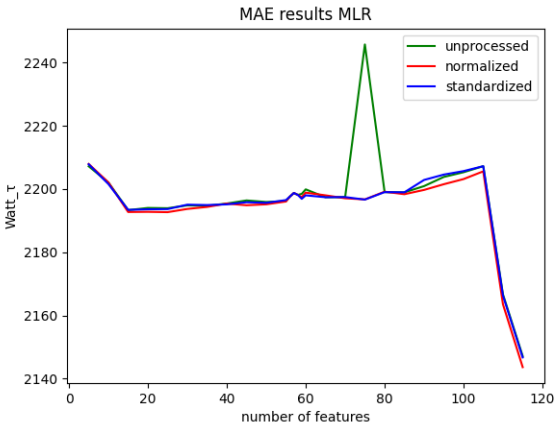


Fig. 7. MAE scores of the Multiple Linear Regression for different data sets

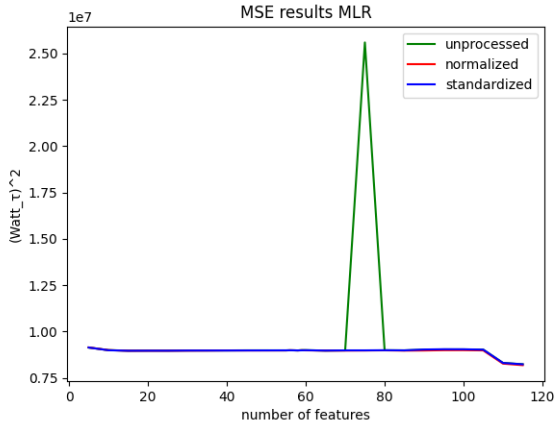


Fig. 8. MSE scores of the Multiple Linear Regression for different data sets

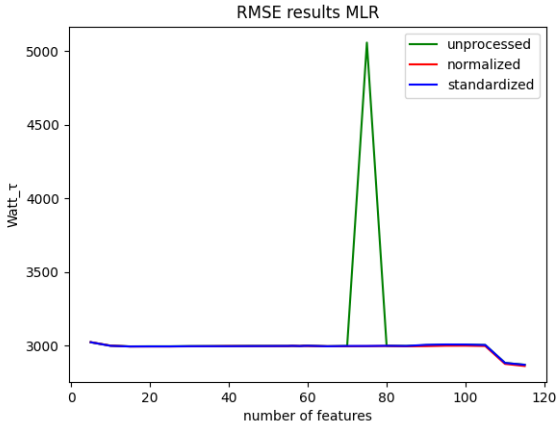


Fig. 9. RMSE scores of the Multiple Linear Regression for different data sets

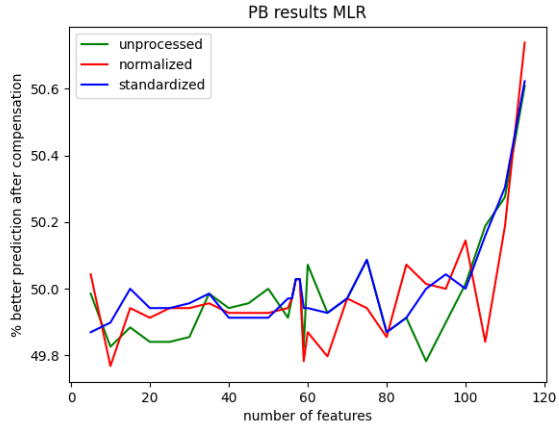


Fig. 10. PB scores of the Multiple Linear Regression for different data sets

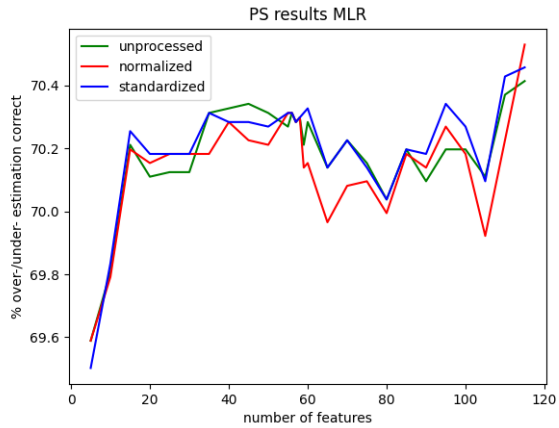


Fig. 11. PS scores of the Multiple Linear Regression for different data sets

A.2 XGBoost

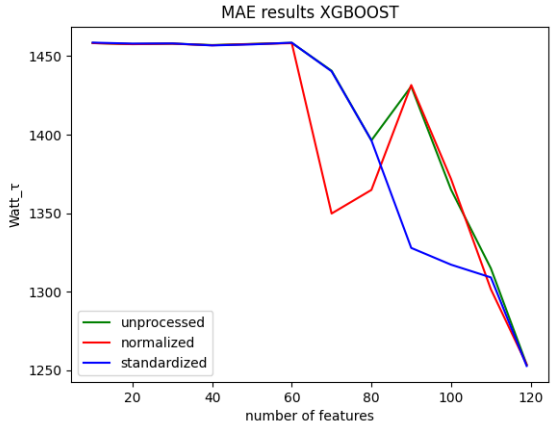


Fig. 12. MAE scores of the XGBoost model for different data sets

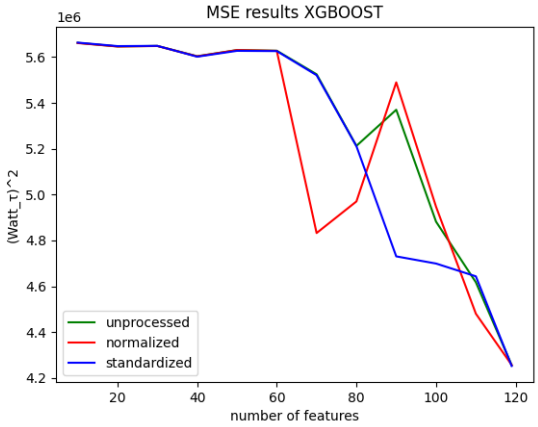


Fig. 13. MSE scores of the XGBoost models for different data sets

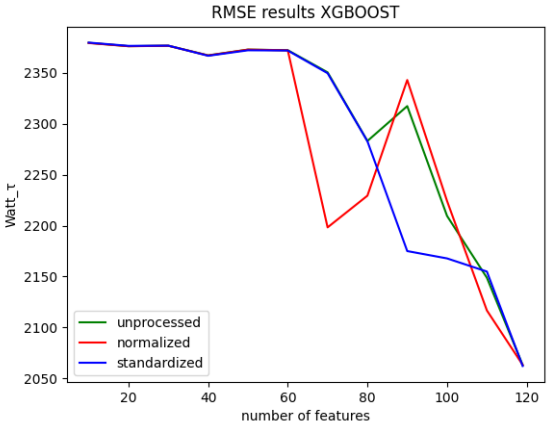


Fig. 14. RMSE scores of the XGBoost models for different data sets

Fig. 15. PB scores of the XGBoost models for different data sets

Fig. 16. PS scores of the XGBoost models for different data sets

B APPENDIX: TABLES NUMERICAL METRIC VALUES

Table 1. Unprocessed MLR results

nr_features	MAE	MSE	RMSE	PB	PS
5	2207.1893447968064	9139200.421457661	3023.111050136541	49.985532407407405	69.58912037037037
10	2202.057016745125	9001390.003355004	3000.2316582815743	49.82638888888889	69.80613425925925
15	2193.3239861405345	8970769.024455681	2995.1242085188524	49.88425925925926	70.21122685185185
20	2194.0150205287086	8975242.668756833	2995.87093659871	49.84085648148148	70.10995370370371
25	2193.9227746100305	8975157.6955126	2995.856754838689	49.84085648148148	70.12442129629629
30	2194.827461283918	8981428.481459748	2996.903148495084	49.855324074074076	70.12442129629629
35	2194.6727990547024	8982943.616800282	2997.1559213361393	49.985532407407405	70.3125
40	2195.411335787018	8983851.489933513	2997.3073732824787	49.942129629629626	70.3269675925926
45	2196.38676602766	8989587.147271326	2998.2640222754444	49.95659722222222	70.34143518518519
50	2195.900522075354	8989029.114934886	2998.1709615922314	50.0	70.3125
55	2196.1680721491375	8986511.876256522	2997.7511364782304	49.91319444444444	70.26909722222221
56	2197.6161461995343	8992509.058102472	2998.751249787563	49.97106481481482	70.3125
57	2198.73968428131	8989216.25869274	2998.202171083988	50.02893518518518	70.28356481481481
58	2198.142598803729	8984143.754857179	2997.356127465867	50.02893518518518	70.2980324074074
59	2198.464125050927	9000853.229488403	3000.1422015445205	49.82638888888889	70.21122685185185
60	2199.8920347554667	8998541.113984123	2999.7568424764236	50.07233796296296	70.28356481481481
65	2197.4017587142484	8976059.598901404	2996.007276176312	49.92766203703704	70.13888888888889
70	2197.4505551421207	8987009.78560775	2997.8341824736985	49.97106481481482	70.22569444444444
75	2245.736724879391	25591462.48620674	5058.800498755287	50.08680555555556	70.15335648148148
80	2199.0184962757085	8994756.317912845	2999.125925651146	49.86979166666667	70.03761574074075
85	2198.931777240617	8987861.51551992	2997.976236650304	49.91319444444444	70.19675925925925
90	2200.9222853304364	9020073.271998998	3003.343681964986	49.78298611111111	70.09548611111111
95	2203.822133376227	9039791.213663897	3006.6245548228826	49.898726851851855	70.19675925925925
100	2205.3092309978183	9041097.142239599	3006.841722179536	50.014467592592595	70.19675925925925
105	2207.2212167938314	9026410.970240066	3004.39860375418	50.18807870370371	70.10995370370371
110	2166.6643278431125	8312814.891956202	2883.195257341445	50.27488425925925	70.37037037037037
115	2147.1196477164362	8240200.495154492	2870.5749415673667	50.607638888888886	70.41377314814815

Table 2. Unprocessed XGBoost results

nr_features	MAE	MSE	RMSE	PB	PS
10	1458.3006813855254	5661801.341843199	2379.4540007832047	67.37557870370371	81.52488425925925
20	1457.6670994820224	5646298.136276899	2376.1940443231692	67.43344907407408	81.52488425925925
30	1457.95509654445	5648467.397942052	2376.650457669796	67.37557870370371	81.42361111111111
40	1457.0433344093083	5603562.796323266	2367.184571663829	66.62326388888889	81.49594907407408
50	1457.7656647700146	5630366.386721055	2372.839309081223	67.47685185185185	81.56828703703704
60	1458.4251996365265	5627562.460696687	2372.2483977645948	66.94155092592592	81.42361111111111
70	1440.6839478735142	5524845.81658502	2350.499056920683	67.43344907407408	81.43807870370371
80	1396.3834801309417	5212449.825721902	2283.079023100581	67.72280092592592	81.78530092592592
90	1430.7434503634636	5370042.421752103	2317.335198401842	66.71006944444444	81.39467592592592
100	1365.006236504631	4881956.290993881	2209.5149447319614	68.28703703703704	82.29166666666666
110	1314.918150873552	4617549.59645109	2148.8484349648975	69.32870370370371	83.52141203703704
119	1253.6623884605006	4256003.2426214535	2063.0082992129364	70.42824074074075	83.70949074074075

Table 3. Standardised MLR results

nr_features	MAE	MSE	RMSE	PB	PS
5	2207.8281001298647	9141208.799389483	3023.443202606836	49.86979166666667	69.50231481481481
10	2201.5696560092747	8994534.650972297	2999.0889701661567	49.898726851851855	69.83506944444444
15	2193.4029333738094	8967007.846646372	2994.49625924735	50.0	70.25462962962963
20	2193.5905323115185	8969543.717343625	2994.9196512333388	49.942129629629626	70.18229166666666
25	2193.6482406326477	8969598.338931926	2994.928770260142	49.942129629629626	70.18229166666666
30	2195.0519859925844	8977464.86530806	2996.2417901945196	49.95659722222222	70.18229166666666
35	2194.9244263008663	8978554.00851262	2996.423536236595	49.985532407407405	70.3125
40	2195.1446470154533	8979489.899877159	2996.5797002377826	49.91319444444444	70.28356481481481
45	2195.8228071593057	8982709.279110132	2997.1168277379734	49.91319444444444	70.28356481481481
50	2195.6268490583648	8985035.1275447	2997.5048169343613	49.91319444444444	70.26909722222221
55	2196.497331851873	8988185.225654695	2998.0302242730468	49.97106481481482	70.3125
56	2197.6161462000687	8992509.058105946	2998.751249788142	49.97106481481482	70.3125
57	2198.73968428123	8989216.258693092	2998.2021710840468	50.02893518518518	70.28356481481481
58	2198.1425988041838	8984143.754858175	2997.3561274660333	50.02893518518518	70.2980324074074
59	2196.8993366600052	8990381.881350258	2998.3965517173106	49.942129629629626	70.3125
60	2198.0023170922823	8995556.485257318	2999.259322775761	49.942129629629626	70.3269675925926
65	2197.401758716053	8976059.598903237	2996.007276176618	49.92766203703704	70.13888888888889
70	2197.450555140983	8987009.78560101	2997.834182472575	49.97106481481482	70.22569444444444
75	2196.656949636202	8986854.922098558	2997.8083531304264	50.08680555555556	70.13888888888889
80	2199.018496273882	8994756.31790901	2999.125925650507	49.86979166666667	70.03761574074075
85	2198.9523869755076	8987906.338092286	2997.9837121125734	49.91319444444444	70.19675925925925
90	2202.8689921925948	9033591.274959126	3005.593331600123	50.0	70.18229166666666
95	2204.566585826147	9049064.53841479	3008.1663083039125	50.04340277777778	70.34143518518519
100	2205.669422318143	9046477.603369253	3007.736292192062	50.0	70.26909722222221
105	2207.227037005617	9026245.352664126	3004.3710411106226	50.159143518518526	70.09548611111111
110	2166.290546871316	8309954.805504876	2882.6992221709284	50.30381944444444	70.42824074074075
115	2146.7163030703264	8236375.921228839	2869.908695625845	50.622106481481474	70.45717592592592

Table 4. Standardised XGBoost results

nr_features	MAE	MSE	RMSE	PB	PS
10	1458.5946353641132	5662942.369480489	2379.693755398053	67.37557870370371	81.52488425925925
20	1457.9647903312705	5647702.986501271	2376.4896352606443	67.43344907407408	81.52488425925925
30	1458.069644497286	5648731.433887947	2376.7060049337083	67.37557870370371	81.42361111111111
40	1456.7994454061254	5601565.749455996	2366.7627150722137	66.66666666666666	81.51041666666666
50	1457.4486785503964	5627072.150002187	2372.145052479335	67.44791666666666	81.56828703703704
60	1458.554195377267	5625974.61509117	2371.9137031290093	66.91261574074075	81.39467592592592
70	1440.3372250060374	5520775.477795762	2349.6330517329216	67.43344907407408	81.43807870370371
80	1396.3529293713932	5211341.215931155	2282.836221880833	67.72280092592592	81.78530092592592
90	1327.9330456627763	4730547.249583187	2174.982126267521	68.63425925925925	82.65335648148148
100	1317.2642263267376	4699301.675385334	2167.78727632241	68.88020833333334	82.88483796296296
110	1309.1628663481263	4643537.793711311	2154.8869561328065	69.140625	82.94270833333334
119	1252.7761803098208	4252353.133102437	2062.123452439848	70.47164351851852	83.73842592592592

Table 5. Normalised MLR results

nr_features	MAE	MSE	RMSE	PB	PS
5	2207.9275885669335	9145173.957741626	3024.098867058024	50.04340277777778	69.58912037037037
10	2202.081826822542	8999931.217253016	2999.988536186933	49.76851851851852	69.79166666666666
15	2192.730155758405	8966071.857713286	2994.339970296173	49.942129629629626	70.19675925925925
20	2192.7819842172125	8967310.303017354	2994.546760866718	49.91319444444444	70.15335648148148
25	2192.6800874621545	8967207.030409768	2994.5295173715967	49.942129629629626	70.18229166666666
30	2193.7004015432217	8972838.729352511	2995.4697009571823	49.942129629629626	70.18229166666666
35	2194.3396258922494	8975953.48232431	2995.989566457852	49.95659722222222	70.18229166666666
40	2195.3918787499824	8983959.177294895	2997.3253372456743	49.92766203703704	70.28356481481481
45	2194.8698731001523	8981504.495535145	2996.9158305723477	49.92766203703704	70.22569444444444
50	2195.166179835974	8981701.910690125	2996.948766777658	49.92766203703704	70.21122685185185
55	2196.0573336802236	8983529.579211662	2997.2536728164437	49.942129629629626	70.3125
56	2197.6161461997294	8992509.058104075	2998.7512497878306	49.97106481481482	70.3125
57	2198.7396842819185	8989216.25869393	2998.2021710841864	50.02893518518518	70.28356481481481
58	2198.142598803988	8984143.754858358	2997.3561274660638	50.02893518518518	70.2980324074074
59	2197.4254412961154	8992457.675262723	2998.7426824025306	49.78298611111111	70.13888888888889
60	2198.8578485313697	8988464.903345736	2998.076867484511	49.86979166666667	70.15335648148148
65	2198.0035703025683	8975918.94664626	2995.983802801053	49.7974537037037	69.96527777777779
70	2197.0749323073514	8975408.450771822	2995.8986048883266	49.97106481481482	70.08101851851852
75	2196.71557944794	8979784.817958135	2996.6289089505453	49.942129629629626	70.09548611111111
80	2199.1578639637764	8988349.27141739	2998.0575830723114	49.855324074074076	69.99421296296296
85	2198.3587051312716	8973797.932515189	2995.6298056527594	50.07233796296296	70.18229166666666
90	2199.715854374913	8980385.255634367	2996.729092800076	50.014467592592595	70.13888888888889
95	2201.5166592018822	8997133.22576368	2999.522166239763	50.0	70.26909722222221
100	2203.144218090152	8998944.267694136	2999.8240394553372	50.14467592592593	70.18229166666666
105	2205.59198840934	8985665.500155838	2997.6099646478087	49.84085648148148	69.921875
110	2163.530772780769	8273600.782340527	2876.386758129116	50.18807870370371	70.22569444444444
115	2143.6440912408793	8190277.112262591	2861.8660192717953	50.73784722222222	70.52951388888889

Table 6. Normalised XGBoost results

nr_features	MAE	MSE	RMSE	PB	PS
10	1458.2773230949376	5660985.154042776	2379.2824872307147	67.33217592592592	81.51041666666666
20	1457.593101668944	5645942.169052414	2376.1191403320695	67.41898148148148	81.52488425925925
30	1457.8982386567188	5648123.621201659	2376.578132778651	67.34664351851852	81.40914351851852
40	1456.8991146494704	5602964.178677729	2367.0581274395713	66.62326388888889	81.49594907407408
50	1457.6770974089034	5630316.242758905	2372.8287428212984	67.47685185185185	81.56828703703704
60	1458.2345660051606	5627096.1800526185	2372.1501175205203	66.94155092592592	81.42361111111111
70	1349.8236797494558	4832306.980002444	2198.250891050074	68.67766203703704	82.74016203703704
80	1364.8410195418373	4970101.741164873	2229.37249941881	68.79340277777779	82.91377314814815
90	1431.5843391812398	5489544.595236548	2342.977719748216	67.28877314814815	82.10358796296296
100	1371.7931862784128	4947094.293859923	2224.2064413763223	68.09895833333334	82.0167824074074
110	1301.6213231607853	4480459.723115112	2116.7096454438697	69.63252314814815	83.34780092592592
119	1253.7284254027024	4256852.903494358	2063.21421657916	70.44270833333334	83.73842592592592