The use of machine learning techniques for predictive maintenance in photo-voltaic systems

By

Jake Connolly

Supervisor

Sarah Sanders

University College London
MSc Computer Science
COMPGC99 Individual Project

Disclaimer: This report is submitted as part requirement for the MSc Computer Science degree at UCL. It is substantially the result of my own work except where explicitly indicated in the text. The report may be freely copied and distributed provided the source is explicitly acknowledged.

September 2018

Abstract

Anthropogenic climate change has highlighted the need to move towards more sustainable sources of energy production. Photovoltaics (PV) make up a large part of global renewable energy capacity, with 98.9 GW being installed in 2017 alone [11]. Despite this, the PV industry is susceptible to shocks by governments, such as unpredictable subsidies, as well as market forces, putting profitability at risk.

Unscheduled maintenance issues have the potential to noticeably effect the power output and therefore profitability of utility scale PV farms. Widespread adoption of SCADA (Supervisory Control and Data Access) systems for data collection provides rich data on numerous aspects of PV farm components. Machine learning techniques have been used with great success for predictive maintenance in industrial settings, limiting the downtime experienced and increasing performance. If this can be applied to predicting the maintenance for PV farms this could reduce the cost of maintenance and maximise the production of energy that can be sold, thereby minimising the risk to profitability resulting from maintenance issues. In this way, the negative impact of external shocks can be more easily managed.

This project aims to investigate the potential for machine learning to predict maintenance issues in PV farms. If this is possible maintenance can be carried out before a system suffers any downtime where it is not producing power.

Data from the Hydes solar farm in the UK produced between 2016 and 2018 was used to trial these techniques. The results were promising, but could be improved upon. Recall values of 0.65 were achieved for a single fault type, predicting ahead by one hour. Whilst this demonstrates that machine learning techniques can predict maintenance issues in PV systems, it is unclear to what extent the methods used in this project can be generalised to other fault types and wider time windows.

Nonetheless, the results from this project provide a solid basis for further exploration of the topic. With more data and more sophisticated techniques, a generalisable, accurate predictive maintenance model for utility scale PV farms should be achievable.

Dedication and acknowledgements

Many thanks to Sarah Sanders for her support and guidance throughout this project. Your confidence in my independence was very much appreciated and your advice always useful.

A further thank you to Darshna Shah and the rest of Elastacloud for providing invaluable suggestions in building the model and providing me with the material that was the ground work for this project. Your enthusiasm for how the project was progressing was always a great motivator.

Finally, a thank you to my cats, Tasha and Lottie, whose undying apathy have been a constant throughout this project.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

As the world moves towards a decarbonised economy, the share of power production from solar has increased dramatically. This increase in many countries has been fuelled by subsidies from governments interested in achieving the sustainability targets that have been set both by themselves and the wider international community [13]. However in doing so capacity has been expanded at a time when typically energy demand is falling, at least in the developed world [1]. In many countries this has led to a slump in wholesale electricity prices [13]. For large scale utilities this has meant that profit margins have become increasingly thin.

As a result, solar utilities are paying more attention to what had previously been considered marginal losses. Three days of downtime results in a loss equivalent to 1% yield loss for solar plants, seriously impacting profitability in the current climate [6]. If this downtime can be reduced, this will have a non negligible impact on the profitability of a plant. If maintenance issues can be predicted in advance this can decrease the downtime of a panel or, ideally, prevent any downtime being experience at all.

In a report analysing the potential impact of machine learning on several industries, predictive maintenance was found to have the greatest impact on the energy sector [23]. Solar panel plants contain a variety of different sensors. A typical 5MW solar PV plant has over 800 sensors which generate about 1 Terabyte of data in a week [3]. This quantity of data makes it difficult to analyse in-house by plant operators but also provides a rich data set on which to use machine learning techniques to predict maintenance issues.

Despite the possibilities for predictive maintenance afforded by machine learning, there has thus far been little research into applying it to the PV domain, despite frequent use in other renewable technologies, notably wind generation [36]. There are multiple reasons for this, chief among these is that the costs involved in PV maintenance are typically only a fraction of the total expenses experienced by a plant. Not only this, but solar modules are predominantly static, a lack of moving parts makes faults significantly less likely. Many common predictive maintenance practises rely on predictable, somewhat linear degradation of components that are correlated to certain metrics measured by monitoring systems, an attribute not present in PV systems. Finally, the vast majority of the UK's large scale solar capacity has been installed in the past 3 years [21]. As the typical warranty for PV systems runs to around 20 years [8], faults are very rare events at such an early stage in their expected usage.

These factors have combined to hitherto limit interest in the area. Yet as companies fleets of solar farms age, faults will become more common. Particularly for large fleets that benefit from economies of scale, rolling out predictive maintenance across gigawatts of installed capacity has the potential to result in significant savings in the future. This situation is true for Quintas Energy, the company that manage the Hydes farm. They manage 2500 MW of solar production over 330 sites [7], so generalisable models for predictive maintenance could prove to be extremely valuable over the long term.

1.2 Aims and Goals

The goals of the project are to deliver:

- An exploration into the data provided from the Hydes solar farm
- Research into various means by which machine learning techniques can be applied to predictive maintenance in the solar farm domain
- Cleaned data that can be used in subsequent models
- A Jupyter notebook detailing the production of a predictive maintenance model, including all necessary cleaning and feature engineering

In addition personal aims over the duration of this project include:

- Improving knowledge of the R programming language
- Improving knowledge of Python programming language and associated technologies
- Improving knowledge of statistical techniques and and algorithms used in machine learning

1.3 Project Overview

The prediction of maintenance issues is the primary concern in this project. The data has been provided in 2 data sets. The first is the SCADA (Supervisory Control and Data Access) data, the second containing the error codes that have been produced by each of the 6 inverters, segmented into 10 modules each. The error codes and their meanings are supplied by the Solar Panel manufacturer, Power Electronics (Appendix A). By analysing both, the project aim is to explore the potential for machine learning techniques to predict the generation of error codes. This can be achieved either as a binary class (for example, will this inverter produce error code 19 in the next two days?), or as a probabilistic measurement of the likelihood of an error code being produced within a certain time frame. The CRISP-DM (Cross-industry standard process for data mining) methodology for data mining [34] was used as the structure for the project. This iterative process provides a sequence of events through which a data mining project progresses. Firstly, as discussed in Section 1.1, the business rational for the project was identified. This elucidates the business success criteria, which in this case is the accurate prediction of maintenance issues via error code generation. The rationale behind this being that this would lead to decreased downtime and increased profitability. Research was also carried out into the typical design of solar farm and machine learning techniques that can be used for predictive maintenance problems. Stage two of CRISP-DM involves understanding the data as it is provided. This was achieved by exploratory plots, assessing the prevalence of each error code as well as missing value quantities and correlations between variables.

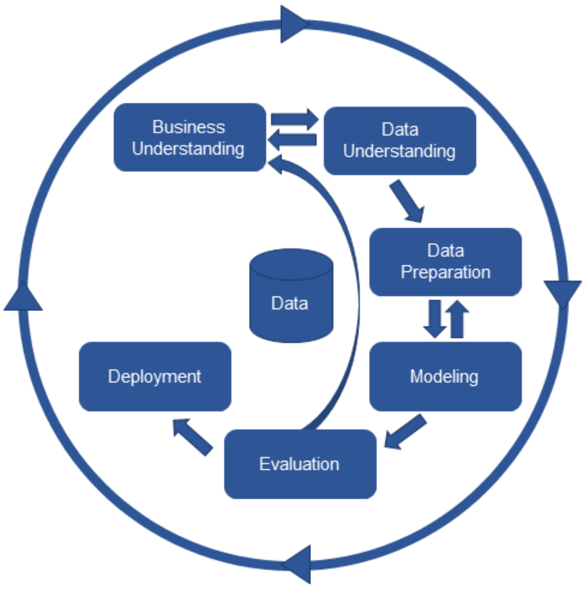


Figure 1.1: Complete CRISP-DM Approach

The data was then prepared for modelling by means of removing missing values by imputation. Feature engineering was used to extract latent data in the data set and features reduction was carried out to reduce the data set. Data from the three separate data sets provided was merged to produce one cohesive dataset that contained all relevant data. Finally, the data interval was altered to be on a more useable scale, from a 15 minute interval to an hour interval.

Models were then built using two separate versions of the final data, one with 20 features and one with 50. The correct accuracy metric is discussed followed by model iteration. Finally an evaluation of the most successful model is performed as well as an interrogation of the feature importance list produced by the model.

The report concludes with a discussion of the project, whether it has satisfied the goals that were set at the start of the project and broaches possible directions in which the results of this project can be taken to improve upon them.

Chapter 2

Literature Review

Solar energy has proven to be a lively field for machine learning in recent years. There has been much research into areas such as irradiance prediction, power output, price point prediction and even predicting soiling events. However there has been little to no specific academic research into the problem of predictive maintenance for solar farms. Therefore, the literature review for this report will focus on the design of solar farms, as domain specific knowledge is an important component any data science project, as well as an investigation into common machine learning techniques used in predictive maintenance in other domains.

2.1 Solar Farm Design

Domain specific knowledge is an important aspect of any data science project. In this specific case, the design and function of utility scale PV farms can provide key insight into the data provided and how features and datasets relate to one another. As well as this, the definitions provided for the error codes require some degree of knowledge of the components present in PV farm design. To this end, research was carried out to explore common design techniques, components and materials used in the industry

PV cells are the basic unit of any PV installation that converts solar energy into electricity. To do this they require a material in which the absorption of a photon raises an electron to higher energy state. Crystalline Silicon wafer solar cells are by far the most common material for this purpose, constituting upwards of 90% of the global PV market [19]. They are rigid and, due to their mass production, the cheapest form of PV cell to produce. Thin film silicon is an alternative material, being more flexible than crystalline Si. There are three main types of thin film solar cells, amorphous silicon (-Si), copper indium gallium selenide (CIGS), and cadmium telluride (CdTe).

Solar farms are constituted of multiple PV modules. The hierarchy of different PV components is as follows. In a typical farm, PV cells are connected in a circuit to form a module. This module is then framed in glass and aluminium so as to protect the underlying cells from materials that might pollute them (dirt or exhaust fumes for example). Multiple modules are then mounted on a stable structure, resulting in a solar panel. A photovoltaic array is the complete power generating unit, consisting of any number of PV modules and mechanically independent panels. Panels must also be linked electrically. Multiple panels connected electrically are referred to as strings. Panels are connected in series to achieve a desired output voltage.

All that is not PV module itself in a PV system is referred to as the Balance of Systems (BoS). Below the key components present in the BoS are detailed.

Mounting System. The mounting system permanently fixes an array to a stationary point. A tracking motor may also be used in the mounting system that enables the modules to be moved to guarantee the maximum sun exposure over the course of the day.

Inverters. Inverters are components that are used to convert direct current (DC), generated by the solar cell into alternating current (AC) which is used by the grid for transmission to where it is needed. Inverters can be responsible for multiple modules or a single micro-inverter can be used fir an individual module.

Storage. Storage is required if the PV system is to exist off grid or if there is a feed in tariff arrangement so that the system charges a battery. As battery technology progresses, the use of industrial scale battery storage is predicted to rise accordingly.

2.2 Predictive Maintenance

The scheduling of maintenance for components or assets can be decided upon in three ways. Most commonly managers will simply follow a run to failure (R2F) program whereby assets are used until they exhibit a failure, at which point an intervention is scheduled. Though common, this is an ineffective method, increasing downtime for assets and resulting in a poor use of labour. A more efficient method is preventive maintenance (PvM). Here, maintenance is carried out on assets according to planned schedule. This is a significant improvement on R2F as most failures are prevented, however still results in regular maintenance downtime, decreasing efficiency [32]

As the quantity of data recorded by industrial machines grows however, predictive maintenance (PdM) has become increasingly feasible [20]. From the data generated by machines, PdF can issues advance warnings for maintenance issues, thereby decreasing downtime from both failure and scheduled maintenance.

The diversity of the scenarios that machine learning based predictive maintenance can be utilised in is such that there is a great deal of information available to gain insight from. There are 4 general strategies that those using machine learning techniques can adopt to attempt to predict maintenance issues, given adequate quality and quantity of data, with each being more or less effective in a given scenario.

The first of these strategies, regression models, can be used to predict the remaining useful runtime (RUL). What constitutes ‘useful life’ is clearly a judgement that must be made on a case by case basis, but this technique typically attempts to estimate the remaining time before a component or asset becomes technically or economically unviable. In a review of RUL estimation methodologies Si et al. [31] highlight that data to come in one of two forms, event data and Condition Monitoring (CM) data. Event data are records of past failures (possibly very scarce) whilst CM data is that which might help provide an estimation of RUL. RUL methods offer potentially very useful and comparatively simple models. That said they can also become very complex. The data best suited to RUL is that has clear failure points, and variables that can sufficiently map the degradation process until the failure point is reached. The benefits of such a system to asset managers are obvious, if one knows how long until an asset or component is predicted to break, then maintenance becomes more efficient.

The second strategy is to frame the data as a classification problem. This is typically achieved by asking the question, will this asset or component fail within a certain time frame. Further, multi-class systems provide the ability to ask this questions for differing time horizons [32]. Classification methodologies remove the constraint placed upon regression RUL for comparatively smooth degradation curves. As a result they are a powerful alternative when the relationships between features are complex.

A third strategy is the use of anomaly detection. This attempts to produce a model for normal function and assign as an anomaly data points that fall outside of this range. Anomaly detection at its most simple can be a using basic statistical methods to determine if a data point falls outside of the standard deviation, or the range around a rolling mean for example.

Machine learning techniques for anomaly detection can be more complex. Density based anomaly detection uses a K-Nearest-Neighbour (KNN) clustering algorithm to identify datapoint are distant from the mean of k data points. The distance can be calculated in a number of different ways, typically using the Eucledian, Manhattan, Minkowski, or Hamming distance. Relative density using a Local Outlier Factor (LOF) can also be used density based anomaly detection [15].

Clustering algorithms are an unsupervised learning technique that assigns data points to clusters based on their distances from cluster’s centroids. Anomalies are identified as those that do not belong to any of the clusters generated by the model. K-Means is a common clustering algorithm.

Support Vector Machine (SVM) might also be used to distinguish anomalies. An SVM creates a discriminative classifier that produces the optimal hyperplane that separates the data into categories [12]. In such a way, anomalies can be identified. The algorithm is typically used for supervised learning (using labelled data), though variations exist that can produce a classifier unlabelled data also, namely OneClassSVM [33].

Survival Analysis is the final technique that is used for predictive maintenance. Similar to

RUL, Survival analysis predicts the probability of failure over time. Survival analysis algorithms can also elucidate the importances of certain variables on predicted probability of failure [35].

Chapter 3

Data

The research and analysis that will be undertaken as part of this project will be based on data from three separate data sets, generated by the Hydes solar farm, operated by Octopus Investments and monitored by Quintas Energy. These include Supervisory Control and Data Access (SCADA) data, Error Code data and PCF data generated by solar array components.

3.1 SCADA

Supervisory Control and Data Access (SCADA) is a software system that directly interfaces with components within an industrial system to perform monitoring tasks so as to gain a more accurate understanding of the health of the component and of the system as a whole. The data generated is processed and made accessible in real time, or in regular intervals. For example. the data generated by the Hydes SCADA system used in this thesis is recorded in 15 minute intervals. This level of regularity allows for more detailed analytical investigation.

The use of a SCADA system in the solar industry is atypical, especially when compared to its pervasive use in other renewable energy generation sources, namely hydropower and wind power [22]. However, as discussed in Section 1.1 , economic forces have decreased profit margins for large scale solar utilities, and thus means of decreasing downtime that were previously perceived as uneconomical are being explored. As well as this, industry groups working in coordination with law makers are pushing for greater uniformity in standards for components [25].

A SCADA system contains four keys features. Sensors on components first receive data, then Conversion units convert this data into digital form. RTUs (Remote Terminal Units) are the most common type of Conversion Unit used to achieve this, owing to their simplicity and wireless

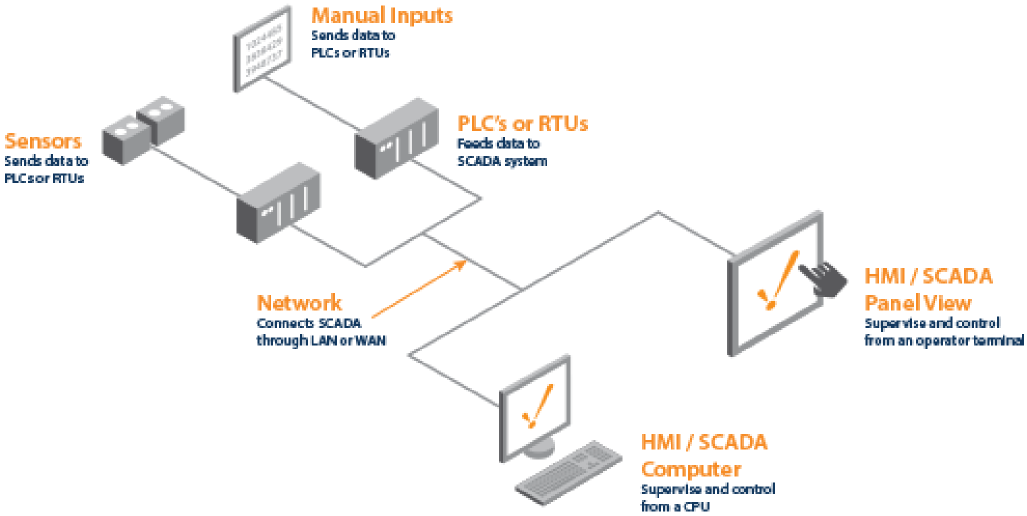

Manual Inputs
Sends data to PLCs or RTUs
PLC's or RTUs
Feeds data to SCADA system
Sensors
Sends data to PLCs or RTUs
Network
Connects SCADA through LAN or WAN
HMI / SCADA Panel View
Supervise and control from an operator terminal
HMI / SCADA Computer
Supervise and control from a CPU

Figure 3.1: SCADA diagram

communication [9]. This digitised data is sent via a communication network (which vary by system) to the Master Unit, a supervisory computer system that handles the Human Machine Interaction (HMI). The information is presented in a graphically intuitive form to the user and is used to compile reports, send notifications or for data mining purposes. In many cases, including the system used in the Hydes, an Operational Historian is employed to update a database of historical time stamped data. It is this data that was used in the analysis for this project.

The SCADA data generated by a solar farm contains information on a wide range of aspects of components across the system, providing a detailed overview of condition and performance. As well as measurements on the internal aspects of the system, solar SCADA systems contain basic external environmental information, such as precipitation and wind speed.

The SCADA data available for analysis in this system can be outlined as follows:

1. Environmental variables, such as precipitation, humidity, irradiance and wind speed and direction

2. Temperature variables at both the string and module level

3. Inverter power variables. These include DC input and AC output as well as reference yields

4. Inverter current variables

5. Inverter voltage variables

6. Inverter energy production and export variables

7. Inverter efficiency variables

8. BoS (Balance of System) losses

The structure of the SCADA data generated by the Hydes system is common-place among solar installations using such a system, and therefore much of the predictive modelling described in this thesis is generalisable to other installations.

3.2 PCF data

PCF data is a mixture of onsite data generated by components and external weather monitoring data used by Quintas Energy to monitor the plant. It contains similar metrics to that of the SCADA data, though not at the same level of granularity. This data was primarily used to compare and check the validity of some of the variables in the SCADA data, though some variables, namely the SOLARGIS variable of satellite irradiation data, provided model improvement independently. Feature selection is discussed in detail in Section 4.5.

3.3 Fault Message data

Fault message data is produced on a module inverter level. As mentioned in Section 2.1, each inverter is responsible for 10 modules. The inverters and associated fault detection system used in the Hydes solar farm is manufactured by Power Electronics. When a fault occurs in a module, this is logged and made known to monitoring facilities. As a result, the faults are logged at sporadic intervals, a fact that must be taken into account when attempting too use SCADA and PCF data to predict the generation of a fault code.

Error Code	Description or Possible Cause
16	IGBT, Gate Drive, cables or Control Boards or Power Board damaged. Power semicondictor internal protection has been activated.
19	Grid current has reached a dangerous level. It's value is above 150% of the inverter rated current. AC power is stronger than DC power.
53	The earth leakage current device has detected an anomaly
55	A fault has been detected in the AC main contactor. Integrated switch doesn't work.
62	Optic Fiber communication fault

Table 3.1: Fault Description Table

There are 61 faults listed in the Freesun HE manual (the inverter model used in the Hydes) (Appendix A). Of these, 26 were exhibited in the data used in this thesis' analysis. Figure 3.2 graphs the frequency of the error codes in the data. It is clear that some error codes constitute a disproportionate amount of the faults generated in the Hydes data and it is these error codes that the model will attempt to predict. Other error codes are generated at frequencies too low to produce reliable predictions. The most frequent five error codes were the targets of prediction, namely error codes 16, 19, 53, 55, and 62. Brief descriptions or possible causes of these error codes are found in Table 3.1.

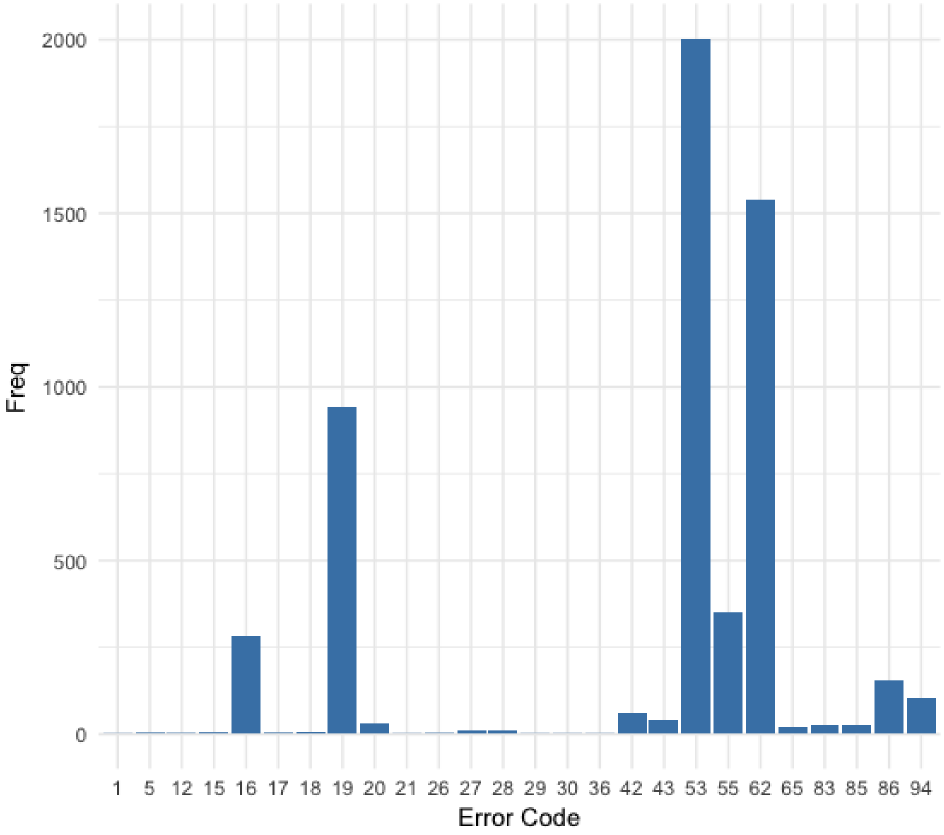


Figure 3.2: Error Code Frequency in Data Set

As domain specific knowledge is of great importance for the production of accurate predictive maintenance models, these descriptions provided key insights that were used to generate meaningful features, as will be detailed in Section 4.5. Indeed it was largely for this reason that the Hydes data was chosen as the trial for predictive maintenance by the client.

Chapter 4

Implementation and Data Preparation

The research and analysis that will be undertaken as part of this project will be based on data from three separate data sets, generated by the Hydes solar farm, operated by Octopus Investments and monitored by Quintas Energy. These include Supervisory Control and Data Access (SCADA) data, Error Code data and PCF data generated by solar array components.

4.1 Requirements

As the nature of this project is the building of a predictive model, rather than software, traditional UML documentation and entity relation diagrams are not an adequate means of assessing requirements criteria. Instead, the below list of requirements MoSCoW process [4].

Key

Prioritisation: M = Must have; S = Should have; C = Could have; W = Won't have

Reference	Requirement	Prioritisation
1	Deliver cleaned data	M
1.2	Research into machine learning techniques for predictive maintenance	M
1.3	A model that provides prediction of error codes to an acceptable accuracy	S
1.4	A model that provides prediction of error codes within a useable time window	S
1.5	Gain background insight into solar farm design	S
1.6	Well documented code in the form of a Jupiter Notebook that details the steps taken to clean and manipulate data and produce models	M
1.7	Begin putting model into production using Azure databricks	C

Table 4.1: Table of Requirements

4.2 Programming Tools and Libraries

The R and Python programming languages were used over the course of the project. R is a statistical programming language widely used in academia. It's popularity has increased with the rise of data science in industry as many of the tools commonly used today were developed in academic settings. It has an extensive array of packages that make it a very good choice for data science projects. Despite Rs popularity, Python and it's ecosystem has become the favoured platform for machine learning [2]. Although a more general programming language not intended exclusively for data science and statistical tasks, Python now boats a rich array of libraries that make it a powerful data science tool. In this project, roughly, exploration and cleaning was carried out in R, as well as some feature engineering. Further feature engineering and model iteration was carried out in Python. Model iteration was chosen to be carried out in Python due to the ease of use of the Scikit-Learn package and the Tensorflow and Keras packages for neural networks (though neural networks are not included in this report).

R packages used:

- **ggplot** Rs the foremost plotting and graphing package used in R
- **XTS** provides easier handling of time based data
- **AnomalyDetection** is Twitter's anomaly detection package

- **lubridate** is another date and time handling package
- **repr** string and binary representations of objects
- **scales** allows manipulation of plot scales
- **corrplot** produces correlation plot
- **entropy** provides entropy value for features
- **caret** Rs core machine learning package
- **gbm** gradient boosting machine package

Python libraries used:

- **numpy** is a well known library for numerical computation
- **pandas** reads data into useable Python data frames
- **matplotlib** primary plotting library for Python
- **scikitlearn** incredibly useful machine learning library
- **random** random number generation
- **xgboost** library supporting extreme gradient boosting machine learning algorithm

4.3 Missing Data

Missing data values were highly prevalent in the SCADA data set. This can be attributed to multiple factors, chief among the diurnal nature of many of the variables with the SCADA and PCF datasets. As PV panels only produce energy during hours in which the sun is shining, the profile of variables that are dependent of energy generation has a steady increase from sunrise until the peak at the noon followed by a steady decline until sunset. Those values present in the dataset during non-producing hours are typically NA due to particularities of the SCADA system (though some variables are logged as zero). These NA values pose a problem in cleaning the data. Data rows with nocturnal NA values might also contain valuable information in other variables, so should not be omitted unless completely necessary.

As well as nocturnal NA values, NA values incurred due to faulty data collection was also present in the dataset. These can produced either by a faulty sensor or incorrect data input from a sensor.

Irrespective of the reason for their appearance, NA values in the dataset must be handled. Methods for doing so depend on the specific dataset. Rows containing NA values can be omitted from the data set entirely, though this results in the loss of data from other variables. Another way is to replace NA values in a variable with a common value, be that 0, the variable mean or some other variable specific value. Although this allows retention of the rows, the information that replaces the NA values will not be useful in most cases and in many cases can severely impact the usefulness of a variable and be detrimental to the accuracy of any model that incorporates that feature.

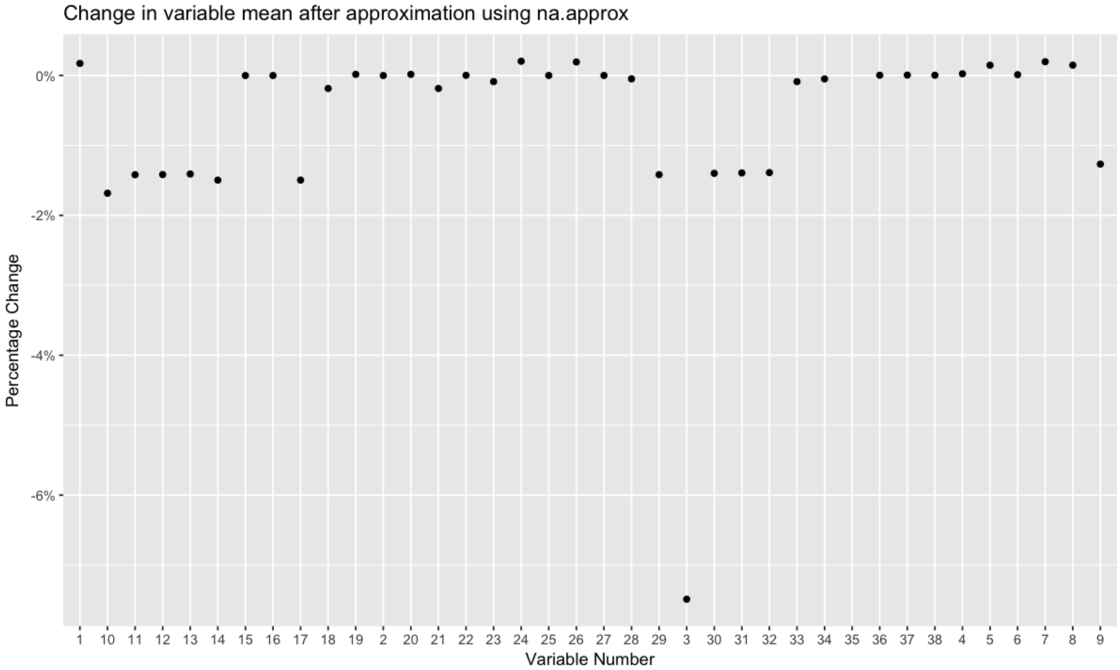


Figure 4.1: Changes in mean after imputation

A final, more sophisticated method of handling missing values involves imputation using the variable data available. There are multiple imputation algorithms available. In this project, Rs off-the-shelf approximation method produced imputed values that did not adversely effect the variable. In order to deduce this, the means and standard deviation of variables containing imputed values were compared before and after imputation. As can be observed in Figure 4.1, imputation had very little effect on the means of these variable, implying that the values generated by the imputation algorithm were reasonable. Other imputation algorithms were also explored, namely MICE [16], however the afore mentioned R na.approx method performed the best out of those trialled.

4.4 Prediction Target

Of the 70 error codes that can be generated by the inverter system, 27 was produced over the duration of the historical dataset. The causation of error codes of course varies, being produced

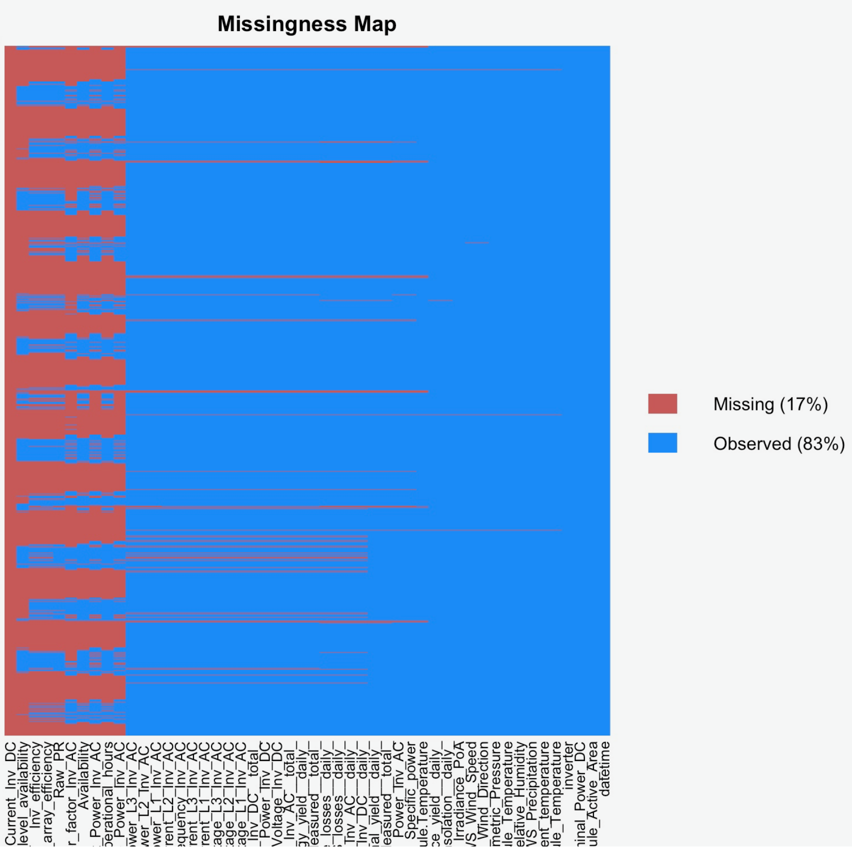


Figure 4.2: Missing data levels by variable

for sometimes very different mechanical and environmental factors. As a result of this limitation, the predictive model will target a single error code, rather than all the error codes that there are instances of in the data set. This will improve model accuracy, as well as allow for more rapid prototyping and better interpretation.

Error code 19 was the choice of prediction target. As was shown in Figure 3.2, it is the third most frequent error code generated in the dataset. Typically, the more instances of the target there are in the model, the greater the accuracy will be. Unlike other error code profiles, error code 19 is also present across inverters, implying a more universal mechanism behind it's generation. Contrast this with the most frequent error code generated, error code 52, which is generated exclusively by a single module in a single inverter. Although understanding why this error code is generated in the inverter is important, a model predicting this error code's generation is unlikely to be generalisable.

Finally, the definition for error code 19 in the inverter manufacturer fault manual defines it as the following: *Starting voltage too low. Grid current has reached a dangerous level. Its value is above 150% of the inverter rated current. AC Power is stronger than DC power.*. This provides an intuitive reasoning behind the error code generation and hints at some variables that might play a role in predicting it. Contrast this to error code 62's definition: *Optic fibre communication fault (HE series)*. This is a much more opaque definition, with none of the SCADA variables being an obvious candidate to link to its prediction.

4.5 Feature Engineering and Selection

Feature Engineering is the process of creating new features from those already present in the data. For example, a variable that reports temperature in Celsius can be used to create a variable that reports temperature in Fahrenheit. The purpose of this is to extract useful features of the data set, that might more accurately describe and predict a target variable.

Feature engineering was carried out in two stages for this project. In the first stage, features were produced across all the variables available in the dataset. The second stage was carried out after a process of feature selection was performed so as to reduce the number of variables in it.

4.5.1 Feature Engineering

The following features were generated before feature reduction.

4.5.1.1 Rolling means

For SCADA variables that still contained a great many null values, the production of rolling twenty four hour means was used to maintain the information captured in those variables. The variables to which this was applicable were `Inv_efficiency`, `Mean_array_efficiency`, `Power_Factor_Inv_AC` and `Raw_PR`. As the data is in 15 minute intervals, the previous 96 values were used to produce a rolling 24 hour mean of the variables in question. Other granularities were explored, however time windows, such as a 4 hour mean contained exclusively null values during nocturnal hours. A twenty four hour mean captures an intuitive statement of the daily values of those variable, allowing them to be incorporated into any subsequent model produced.

4.5.1.2 DC_AC_diff

As stated in Section 4.4, the definition for provided by the inverter manufacture for error code 19 generation is as follows: *Starting voltage too low. Grid current has reached a dangerous level. Its value is above 150% of the inverter rated current. AC Power is stronger than DC power*. To attempt to capture information relevant to this description, a feature reporting the difference in the AC power (alternating current passed to the grid) and DC power (direct current produced by the solar panels). This was achieved by simply subtracting the DC power for an inverter from the AC power for each 15 minute interval.

4.5.1.3 PCF Data

Although not strictly feature engineering, the PCF data was added at this juncture in the process.

4.5.2 Feature Selection

After initial feature engineering, the number of variables present in the dataset was reduced. Reducing the dimensions of a dataset is important, as excess variables add unnecessary complexity to a model and can prevent information that is relevant to the prediction of the target.

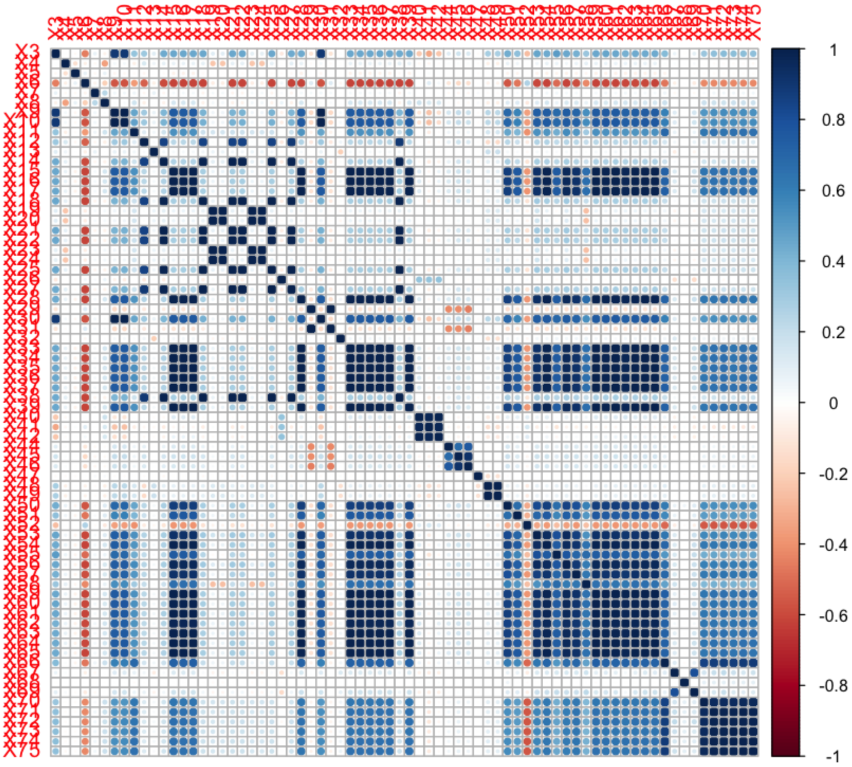


Figure 4.3: Correlation plot of variables

The first step in this process is to analyse the correlation between the variables. Correlation is a value between 1 and -1 that describes the extent to which changes in one variable are mirrored in another. A correlation of 1 would mean that as one variable increases so does the other by a proportionate amount. A correlation of -1 would mean the opposite, as one variable increases, the other variable would decrease by a proportionate amount. A correlation plot represents this graphically (Figure 4.2).

This information is used to deduce which variables contain similar information to one another. Where this is the case, variables can be removed to reduce the complexity of the data set. For each variable a list of variables that showed a correlation of greater than 0.9 was produced. Via a

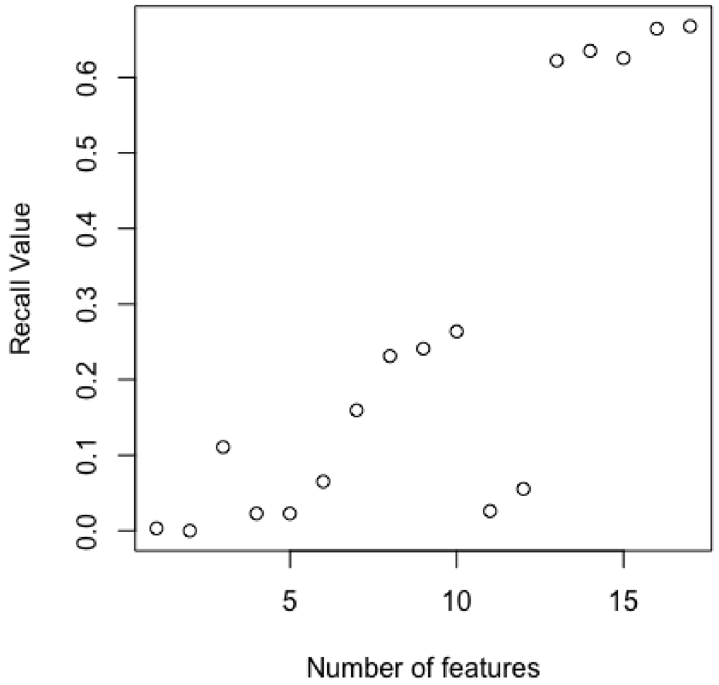


Figure 4.4: Effect of feature addition on recall value

largely heuristic process of elimination, variables were chosen to incorporated into the model that captured the maximum amount of information, whilst others were dropped from the model. This heuristic process also incorporated the feature importance produced by earlier model iterations, which many algorithms can produce (in this case a gradient boosting machine). This process reduced the number of variables in the dataset from 70 to 20.

An entropy analysis was then carried out to determine whether the number of variables could be reduced further. Conditional entropy provides a measurement of how much more certain we are able to be about our target variable if we know the value of a particular variable. By ranking variables by their entropy (represented by Mutual Information) we can add each variable starting with the most important to a basic model and measure the effect on the accuracy of the model. In Figure 4.3 we can see that to achieve the maximum accuracy, all 20 variables were required, and therefore no further feature reduction can be carried out without harming the accuracy of the models they are used in.

4.5.3 Feature Engineering Continued

Two data sets then underwent further transformations. These data sets are the reduced feature dataset consisting 20 features (hereinafter referred to as 20F) and a dataset consisting of all 50 features of the original data. Although the process discussed in Section 4.5.2 appeared to show that reducing the dimensions of the dataset would increase the accuracy of a model, it is necessary to retain a larger feature dataset lest this not be the case. A 50 feature dataset (hereinafter referred to as 50F) was used instead of the 70 feature full dataset, as the correlation list and Gradient Boosting Machine feature importance showed that the PCF data that constitute the remaining variables were either negligibly important or their information is captured by a SCADA variable.

The 20F and 50F datasets were then manipulated to produce the following variables.

4.5.3.1 Hourly means, maximums and minimums

To capture the temporal structure of the dataset is important to create features that in some way represent the changes over a period of time. To that end, the data received in 15 minute intervals was rolled up to produce hourly means, maximums and minimums for all variables measured variables. In this way each entry represented data over the course of an hour.

4.5.3.2 Modules_affected

In order to capture all the data available in the error code dataset, a feature was produced that stated the number of modules affected in each inverter for a given hour. This shows something of the 'magnitude' of the error code event, the number of modules a single inverter is responsible for that demonstrate the error code, rather than simply a binary value.

4.5.3.3 Target Variable Lag - Error_Code_nxt_hour

Finally, the target variable is manipulated. The error data that is to be combined with the SCADA data contains a variable named `Error_Code`, a binary value that states whether an error code has been produced in a given inverter during this hour. This is inadequate for the models for two reasons. Firstly, the engineered feature `Modules_affected` directly predicts this (if at least one module in an inverter generates an error code, then by definition an error code has been generated by that inverter). This produces false levels of accuracy as a predictive model will always be able to correctly predict those values.

Secondly, a predictive model that only provides the solar farm operator a prediction for whether an error code will be produced in that very same hour is of little use to the end user of the system, profitability increases from decreased maintenance downtime would be negligible. As a result the error code variable is lagged back an hour to essentially predict whether an error code will be generated in the subsequent hour given the values of the current hours variable readings.

A one hour prediction window is also quite possibly of small use to the farm operator, a matter discussed at greater length in Section 6.1, but allows this thesis to demonstrate clearly the potential for a predictive model. Other predictive time windows were experimented with, with varying degrees of success, however one hour provided an adequate mix of both accuracy and foresight.

50F and 20F are then combined with the error code dataset to be used as the final data during model iteration.

Chapter 5

Model Iteration

A series of models were implemented using the two datasets, 20F and 50F produced as described in Section 4. These models can be divided into two separate categories. Two main predictive maintenance methods will be utilised in this thesis, namely anomaly detection and classification.

Each model, initially using standard parameters, was used to predict the target variable `Error_Code_nxt_hour`. The data was split into training and testing parts. The model is trained on data that includes 75% of the available data, where 75% of the positive target variable cases are present. The remaining 25% is used for testing the models accuracy.

5.1 Accuracy Metrics

The method by which the model accuracy is determined is important, especially when the data used to train and test the model has very high class imbalance. The data used in this thesis has a very high class imbalance, only 0.38% of the target variable were positive class. Commonly, accuracy is the metric by which a model is judged. This provides a measurement of how many data points were accurately predicted by the model. However when class imbalance is high, this is a poor metric to use. For example, in the data used here a model might predict a negative class for all datapoint and still achieve an accuracy score of over 99%.

Recall, precision and F1 scores are used to mitigate this problem. Precision provides a measurement of the ratio between false positive and true positive values. This is a useful metric when the consequences of a false positive are worse than that of a false negative. A false positive in the data used for this analysis would correlate to the model predicting an error code when there it should not have. In this domain, this is not a great problem, as the this would occur so infrequently that the impact that unnecessary checks will incur on the implant would be negligible. As a result, precision is a useful to take into account, though not very useful as a means of deducing a model's predictive efficacy.

$$Precision = \frac{True \quad Positive}{True \quad Positive \quad + \quad False \quad Positive} \tag{5.1}$$

Recall returns a value that represents the number of positive results, in this case error codes, that are correctly predicted by the model. This is a significantly more useful metric, as it is far more important for the model to correctly identify when an error code will be generated. If a positive result is missed it could lead to more significant maintenance issues at a later date. For this reason recall will be the primary metric of interest for judging model effectiveness.

$$Recall = \frac{True \quad Positive}{True \quad Positive \quad + \quad False \quad Negative} \tag{5.2}$$

Finally, an F1 metric provides value that is intermediate from recall and precision. This too is a useful metric, as it captures something about the strictness of the model. This will be useful when comparing models, as if one has a lower recall but higher F1 it implies that parameters and can be manipulated to encourage a less strict prediction, making the model more inclined to predict the positive value. A cost matrix might be a way of achieving this.

$$F1 = 2\frac{Precision \quad x \quad Recall}{Precision \quad + \quad Recall} \tag{5.3}$$

5.2 Anomaly Detection

The sole anomaly detection algorithm investigated was OneClassSVM. As discussed in Section 2.2, the purpose of anomaly detection algorithms is to provide representational model of a system, whereby any reading that falls outside of this normal functioning is deemed an anomaly. OneClassSVM is an unsupervised method, meaning the algorithm does not make use of labelled data.

The advantages of attempting to use an anomaly detection method for this problem is that, as the number of positive cases are very small, an algorithm that is able to identify novel/anomalous behaviour without seeing any past examples of that behaviour is very useful. Not only this but it would scale to include other error codes with greater ease than traditional classification algorithms.

Support Vector Machines (SVMs) are a type of algorithm that is used in both classification and regression problem. Data points are separated in space in such a way that a plane can be used to separate data points that fall into one category from those that fall into another. This plane is referred to as a decision boundary, new data points are mapped to this model and classified according to which side of the decision plane they fall on. Decision planes can be non-linear by projecting the data into higher dimensions, in this way a hyperplane can be used to separate data points that were non-separable in linear space [30].

OneClassSVM is a version of this classification approach first posited by [27] that separates all the datapoint from the origin and maximises the the distance from this hyperplane. This results in a function that presents regions in input space where the probability density of data is high. This binary function returns a '1' value for data points inside these regions and '-1' outside. This function is stated below.

$$
\begin{aligned}
\min_{w,b,\xi_i} \frac{\|w\|^2}{2} + \frac{1}{Vn}\sum_{i=1}^{n}\xi_i - \rho \\
subject \quad to: \\
y_i\left(w^T\phi(x_i)+b\right) \geq \rho - \xi_i \qquad for \quad all \quad i = 1,...,n \\
\xi_i \geq 0 \quad for \quad all \quad i = 1,...,n
\end{aligned}
\tag{5.4}
$$

The defining parameter here is V, this sets the upper of the fraction of outliers and a lower bound on number of training examples used as a support vector. The V value used in the OneClassSVM model in this thesis was 0.0038 (rounded to 2 s.f). This was arrived at by calculating the fraction of the target class in the data set.

5.3 Classification

Classification algorithms separate the data into data points that will generate an error code in a given time frame and those that will not. The time period to be predicted in this problem is the subsequent hour. Classification algorithms formed the majority of the algorithms explored in this project.

5.3.1 Decision Tree

Decision tree is a powerful, intuitive supervised learning algorithm used for classification and regression. Put simply, each node in a tree correlates to a feature, each branch and decision based on the value of that node, and each leaf is an outcome of that decision. Decision trees split the data into subsets, based on values of the variables in the dataset. If a variable always predicts the same class at a given value it is referred to as pure and the tree stops. However if a variable is not pure the tree moves onto another attribute where another split occurs. A greedy algorithm recursively moves down the tree until exhausted or until the maximum depth parameter is reached, if this was incorporated. There are multiple algorithms used to build decision trees, the two most common are investigated in this thesis. They are CART (Classification And Regression Trees) that uses a Gini Index metric, and ID3 (Iterative Dichotomiser 3) that uses entropy function as a metric.

5.3.1.1 ID3

To decide which variable to use as the root of the decision tree, and which variable should be used in subsequent splits, the ID3 algorithm uses that which has the highest information gain. To do this, the entropy must be determined. Entropy is a measure of the uncertainty about the data. If all classes are the same the entropy will be 0, if classes are split evenly, entropy is maximum at 1. Information Gain (IG) is a measure of the difference in entropy before and after a dataset S is split on variable A [10]. The IG is determined for each variable and the one with highest is set as the root node. The variable with the next highest IG is split at the next level of the tree if the root variable is not pure.

5.3.1.2 CART

In the CART algorithm the Gini score is used to evaluate the splits of the dataset. A Gini score reflects how well mixed the classes are by a split. If classes are split evenly this gives a Gini score of 0, a split in half gives the maximum of 0.5. The process is very similar to the ID3 algorithm except the difference in Gini scores before and after a variable is added are compared, not entropy. The variable resulting in the highest Gini score difference becomes the root with the second highest added after and so on.

5.3.2 Random Forest

Random forest is an algorithm that is similar to decision tree, but instead of a single tree creates an ensemble of separate trees created using subsets of the original data. The values are then averages together. In this way, instead of choosing to split by the most important feature in the dataset, is does so by the most important feature in this random subset [14]. This mitigates the problem of overfitting that decision trees can suffer from. The disadvantage is that they are less interpretable than decision tree, as they do not consist of a single tree, but rather is a amalgamation of many.

5.3.3 XGBoost

XGBoost (Extreme Gradient Boosting) implements the gradient boosting decision tree algorithm [18]. Boosting is a so-called ensemble method, whereby weak models are added to one another to minimise the errors produced by each model. Models are added until no further improvement can be made. Gradient boosting is an extension of this where new models are built that predict the residuals of previous models before being added together to provide a final prediction. The error loss is minimised using a gradient decent method, hence ‘gradient boosting’. XGBoost has proven to be a very effective algorithm across many problem types and has the advantage of being exceptionally fast to run. It is an improvement on the Decision Tree algorithm described above as

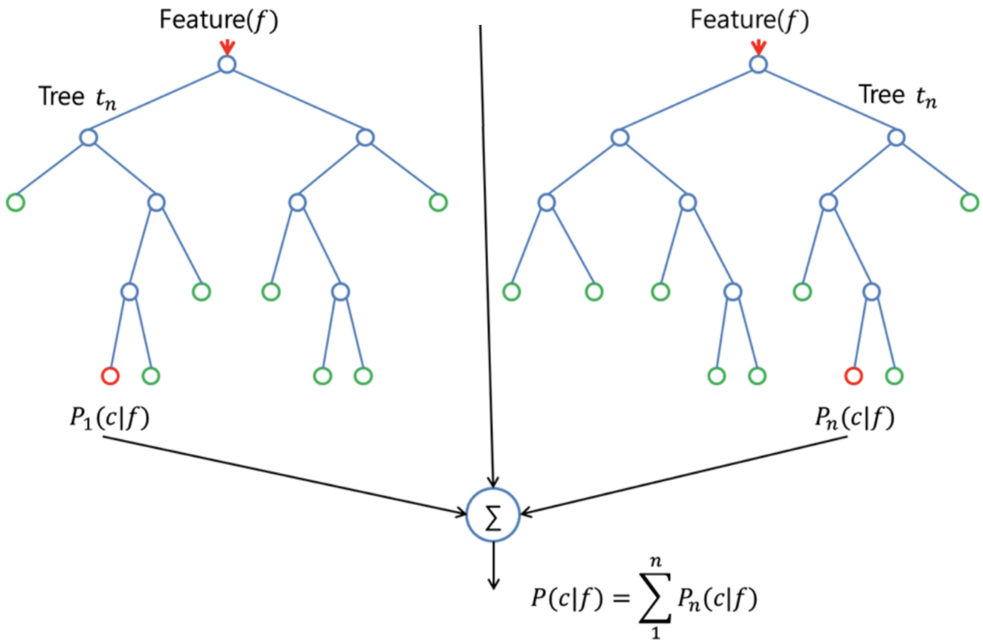


Figure 5.1: Random Forest Diagram

these tend to overfit on trained data, ensemble methods mitigate this problem by building many separate, smaller trees.

As this algorithm performed well in initial trials, a grid search was carried out to tune the model's hyper-parameters, focusing on the `min_child_weight` and `max_depth` parameters. Evidence from other applications of XGBoost suggested that these parameters were good candidates to alter for improved performance [5]. Tuning these parameters did not have an effect on the recall value of the model.

5.3.4 K-Nearest Neighbours Classifier

K-Nearest-Neighbours (KNN) classification is a non parametric, supervised learning algorithm that attempts to predict the class of a datapoint based on the K nearest neighbours of that datapoint. The class value is determined by the equation

$$y = \frac{1}{K} \sum_{i=1}^{K} y_i \tag{5.5}$$

in which K is the number of neighbours, y is the prediction and yi is the class value at each neighbouring data point. This distance can be determined by multiple methods, typically the euclidean distance is used. As the algorithm is non-parametric, that is, there no parameters to be added apart from the k value, KNN is a simple and intuitive algorithm to use in many cases. However, it is computationally intensive and largely unable to capture more complex relationships in the data.

Another common classification algorithm is a Support Vector Machine (SVM), much like unsupervised OneClassSVM, this supervised algorithm separates the data using a non-linear hyperplane. Whilst this algorithm has attributes that make it a promising candidate for this classification problem, it is computationally intensive. The limitations of the machine on which these models were trained (2.3 GHz Intel Core i5 processor), made training an SVM an unfeasibly long endeavour, despite methods used to try to decrease the time.

Chapter 6

Results and Discussion

6.1 Model predictions

The recall and F1 scores for each model is presented in Figure 6.1 and Table 6.1. As shown in these figures, the 20F dataset produced the best recall scores, scoring 0.48 on average. In contrast the 50F models scored on average 0.46 recall. After rounding to 2 decimals places, the average F1 score for both 20F and 50F models was 0.52.

Interpretation is made somewhat more difficult by the difference in Recall and F1 Score. The model with the highest Recall score is a the 20F dataset decision tree, using the entropy based ID3 algorithm. However, the model with highest F1 value is the 50F XGBoost algorithm. The high F1, as well as fairly high recall, shows that the 50F XGBoost model has very high precision (indeed it was 88%). This implies that the model parameters can certainly be altered in any further iterations to make it more inclined to predict the positive value. Should altering parameters not help increase the recall value, introducing a cost matrix might affect the outcome also. For this reason, the 50F XGBoost model is that which is suggested by this report as a basis for predictive maintenance for error code 19 prediction in the Hyde's solar farm.

The feature importance produced by the 50F XGBoost model is largely an intuitive reflection of what would be expected given the manipulations carried out on data and the definition of the target variable. Figure 6.2 presents the 20 variables with the greatest importance in model. The method the importance is derived from uses the number of times a feature is used to split the data across all trees. interpreting these results, the inverter number is likely to have been given an artificially inflated importance as having only 6 values means that choosing any of them appear to have a large effect on the decision tree. The `modules_affected` variable is important as it provides an snapshot of the previous hour's error code 19 generation. Error

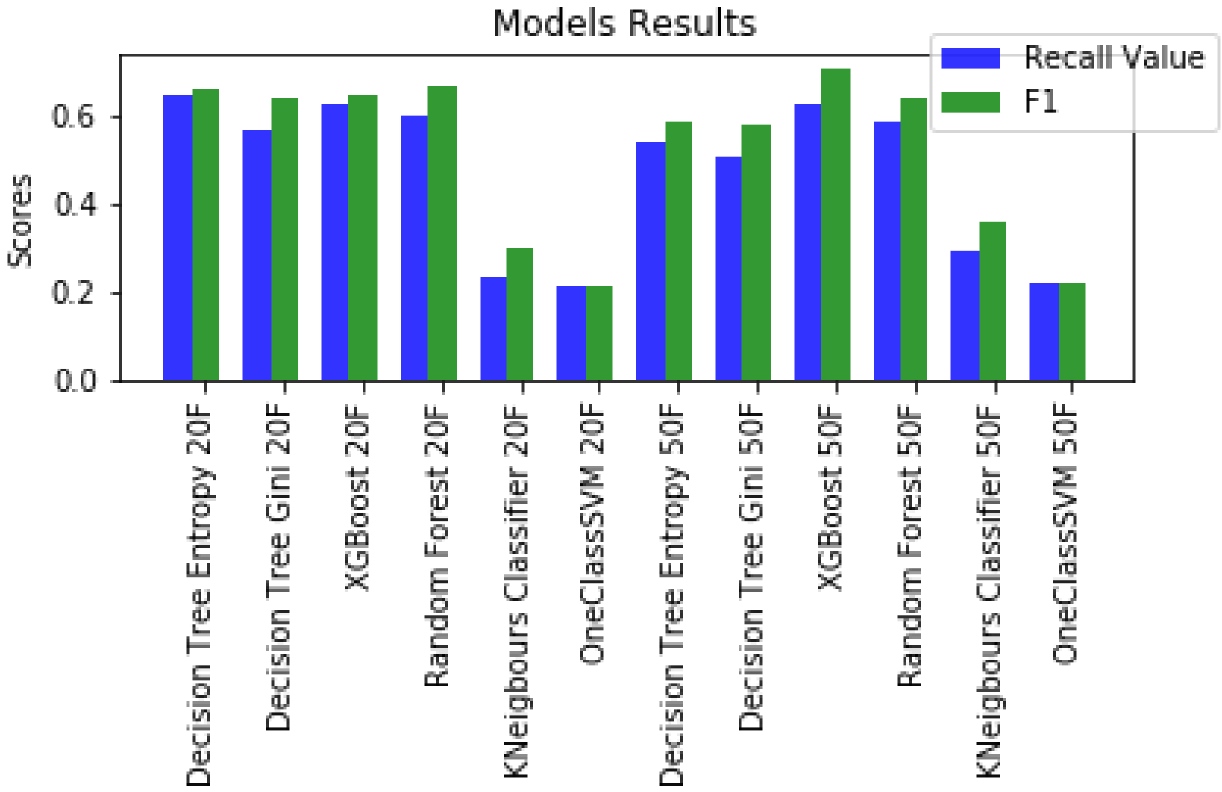


Figure 6.1: Model Recall and F1 results

codes tend to be triggered in bursts, so knowledge of the previous hour is a strong predictor. `String_5.01.01_Module.Temperature, amin` makes an intuitive sense. During the exploration of the data, it was clear that temperature and irradiance variables correlated highly with instances of error code 19 generation. Similarly, `Irradiance_PoA` mean is an important feature for the same reasons. This can be explained somewhat by the correlation between irradiance and temperature. As the solar cells are exposed to more of the suns energy, the irradiance, both ambient and solar module temperature increase.

Model	Recall Score	F1 Score
Decision Tree Entropy 20F	0.65	0.66
Decision Tree Gini 20F	0.57	0.64
XGBoost 20F	0.63	0.65
Random Forest 20F	0.60	0.67
KNeigbours Classifier 20F	0.23	0.3
OneClassSVM 20F	0.20	0.20
Decision Tree Entropy 50F	0.54	0.59
Decision Tree Gini 50F	0.51	0.58
XGBoost 50F	0.63	0.71
Random Forest 50F	0.59	0.64
KNeigbours Classifier 20F	0.23	0.3
OneClassSVM 50F	0.21	0.21

Table 6.1: Model Performance

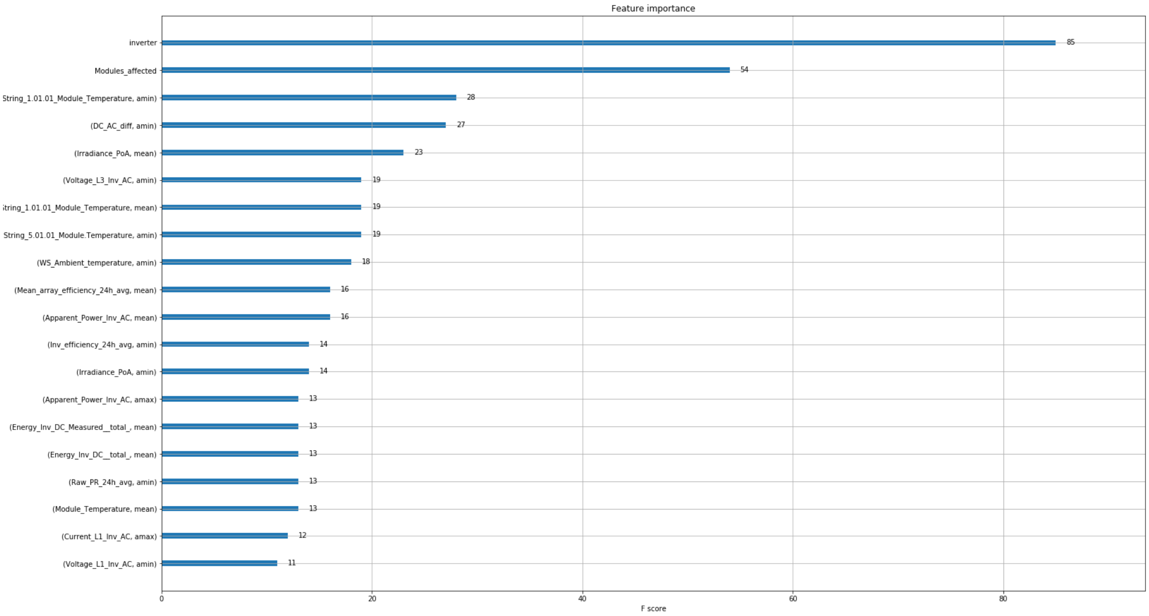


Figure 6.2: XGBoost Feature Importance

CHAPTER 7

CONCLUSION

The stated aim of this project was to explore the potential for machine learning techniques to produce predictive maintenance models for the Hydes solar farm, with a view to scale a successful and useful model for use in all Quintas Energy managed solar farms. This goal has been achieved with mixed success, though the results contained in this report are certainly promising. It has been established that by framing the problem of predicting error code generation in terms of classification, traditional classifications algorithms can be deployed to produce predictions with fairly good accuracy.

There are a number of areas, of course where the model produced in this project must be refined and improved upon before being deployed on a commercial level. Firstly, error code 19 is but one of a multitude of error codes that the HE inverter can produce. Predictions of many of these will no doubt remain unfeasible, due to the sparsity of their occurrences in the dataset. As more data becomes available, both at the Hydes and other solar farms managed by Quintas Energy, the feasibility may be change. Secondly, the time window in which a prediction can be made must be increased to be commercially applicable. Though not necessarily the purpose of this project, future work must make a trade off between the accuracy of the predictions and how far into the future they make these predictions.

7.1 Future Work

Future work should focus initially on the production of the current model using big data technologies and platforms. Initially with will involve the use of Apache Spark to carry out the requisite data cleaning and manipulation before data is passed to the model. This can be achieved by using Microsoft Azure's Databricks platform, that streamlines model iteration and data cleaning in the cloud.

Further to this, there are multiple techniques that might be attempted in order to produce a more accurate model. Above all incorporating more data from the Quintas Energy fleet will undoubtedly improve the ability of any model to make predictions. The size of this data will necessitate moving to platforms that are better able to work with Big Data, using technologies such as Apache Spark described above.

7.1.1 Other Machine Learning Techniques

Some machine learning techniques not adequately explored in this report are discussed below.

7.1.1.1 Neural Networks

Artificial Neural Networks (ANNs) are a class of machine learning that uses methods that are analogous to the way biological neurones interact [29]. Artificial neurones containing values are arranged in layers, with an input layer of features at the start and an output layer of predictions at the end. In between these layers are one or hidden layers that store the outputs after the previous layers neurones are fed through a function. The learning element occurs via back propagation that compares the predictions arrived at by the network with true values. This modifies the weights that connect neurones between layers until an optimum is produced. The maximisation metric used during this back propagation is accuracy.

As discussed in Section 5.1, accuracy is a poor metric to use for assessing predictions made on a dataset that suffers from severe class imbalance. If a neural network attempts to maximise the accuracy for class predictions, the recall value for the positive class is poor (in exploratory models used in this project recalls of 48% were achieved). That said, neural networks are incredibly powerful and there are ways in which the class imbalance problem can be circumvented.

Although infrequently used, neural networks that use F1 as a maximisation metric are possible [26]. However the process is not trivial, it is not supported by easy to use packages such as Tensorflow and Keras that make building neural networks a comparatively easy task. Whilst it may be promising, there are other means by which the power of neural networks can be leveraged for this problem.

As class imbalance is the reason that neural networks are not effective for the data set as it currently is, fixing this problem will allow a traditional Neural Network to perform better. SMOTE (Synthetic Minority Oversampling Technique) is a method of 'oversampling' minority classes by creating synthetic instances of them by using K-Nearest-Neighbours in feature space [17]. As well as 'upsampling' the minority class, the dominant negative class will have to be downsampled, a process which must be undertaken with the objective of maintaining the maximum variance of negative class variable values. Correctly selecting the majority class data points will be a slow process but this technique, combined with traditional neural network methods shows a great deal of potential and is certainly an area that should be explored in any further work on the topic.

A further suggestion for using neural networks for this problem is anomaly detection using a LSTM (Long Short Term Memory) Neural Network. The temporal quality of the Hydes data is captured by extracting features of hourly means, maximums and minimums in this projects case. However LSTM Neural Networks are good at learning patterns over time periods due to their longer memory. This can be used to model normal time series, which in turn can be used for anomaly detection [24]. The nature of the data used in this project means that it is still doubtful whether anomaly detection methods will useful, though it is an avenue that bares exploring due to the scalability of an anomaly detection approach.

7.1.1.2 Dynamic Time Warping

A separate means of capturing time series data, aside from extracting time based means, maximums ect, is to use a KNN approach. In this approach, for a given time series example, find the most similar time series in the training set and use its corresponding output as the prediction [28]. This requires finding a measure of similarity. Dynamic time warping finds the optimal non-linear alignment between two time series in the data. After the similarity between two time series can be deduced, KNN can be used to find the most similar time series in the data set and a prediction made based on the value of the target variable in this time series. This again is a promising method that should be explored. It should be noted though that it is computationally very expensive.

APPENDIX A

Bellow the manual for the Freesun HE series inverters produced by Power Electronics is provided

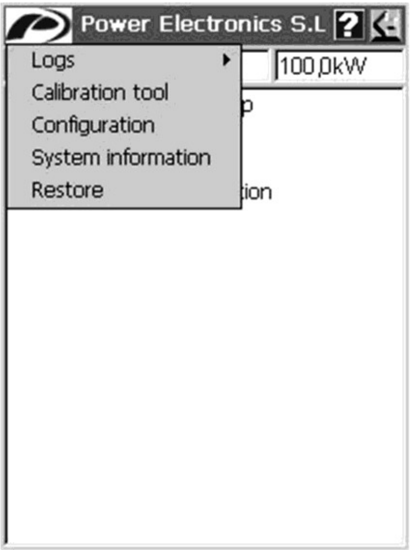
A.1 Inverter Fault Manual

8. FAULT MESSAGES

When a fault develops, the FREESUN will stop, showing on the display the fault occurred in the status line (second line above), the values of some parameters will "frozen" temporarily and this information will be stored in internal files which are fault registers, to be inspected later, after resetting the system. This function is a valuable source of information for both technicians and end user. Thus, it is possible to know what was happening in the system when it failed.

Moreover, the FAULT LED will be on and the fault message will remain until the fault is sorted out and the inverter is rearmed.

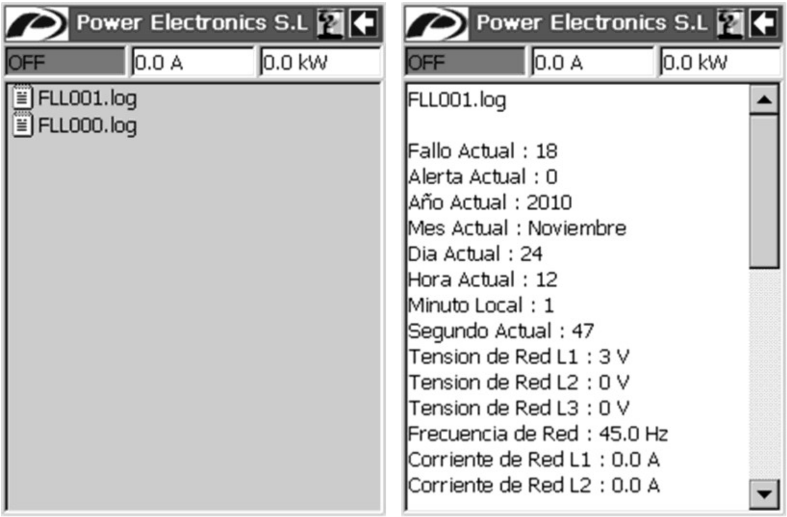
Pressing '**P**' key, the system menu will be displayed and the fault register will be accessible through the Logs→View option. The Delete option is password protected and is only for technical service personnel. To exit the fault register visualization screen and go back to the home screen, tap on the button '**←**'. To return to the screen displayed before entering the fault register home screen, tap on the button '**←**', or select the 'Logs' option tapping on it from the system menu.



FSITC0031AI

Figure8.1 System Menu

All faults are stored in separate files. The user can access these files by tapping on the file name. Then, all information concerning to this fault will be displayed.



```
FLL001.log

Fallo Actual : 18
Alerta Actual : 0
Año Actual : 2010
Mes Actual : Noviembre
Dia Actual : 24
Hora Actual : 12
Minuto Local : 1
Segundo Actual : 47
Tension de Red L1 : 3 V
Tension de Red L2 : 0 V
Tension de Red L3 : 0 V
Frecuencia de Red : 45.0 Hz
Corriente de Red L1 : 0.0 A
Corriente de Red L2 : 0.0 A
```

FSITC0012AE

Figure 8.2 Logs screen information

8.1. Description of Fault List - Troubleshooting

DISPLAY	DESCRIPTION OR POSSIBLE CAUSE	ACTIONS
F00 NO FAULT	The inverter is operative. There is no fault.	None
F01 HI. DC VOLT. HW	*Over Voltage Vdc H/W* High DC voltage of solar panels. HW has detected that panels DC voltage has reached a dangerous level >900Vdc.	Check de real DC voltage with an external voltmeter and the displayed voltage in the inverter module. Replace the control board. Replace the power board.
F02 HI. DC VOLT. SW	*Over Voltage Vdc S/W* High DC voltage of solar panels. SW has detected that panels DC voltage has reached a dangerous level >900Vdc.	Check de real DC voltage with an external voltmeter and the displayed voltage in the inverter. Replace the control board. Replace the power board.
F03 HI DC CURR. HW	*Over Current H/W* HW has detected that panels DC current has reached a dangerous level 200% above the inverter rated current Damaged control board (HE series) Cable of the Current DC Transformer damaged. Current DC transformer damaged.	Check the DC current transformer wiring For HE Series replace the control board Replace the power board. Replace the DC current transformer
F04 HI DC CURR. SW	*Over Current S/W* SW has detected that panels DC current has reached a dangerous level 200% above the inverter rated current Cable of the Current DC Transformer damaged. Current DC transformer damaged. Damaged control board or power board	Check the DC current transformer wiring Replace the control board Replace the power board. Replace the DC current transformer
F05 INVERSE DC POLARITY	*Inverse DC polarity connection*	During commissioning check that the DC+ and DC- input terminals are correctly installed.
	Motorized Circuit breaker failure The DC bus voltage did not reach the set value within the time established. Therefore the MCB or fuses are locked or faulty.	Reset the MCB by using the manual KEY and turn it a complete turn of 360º, finally close the gate to leave it in manual. Check that the MCB voltage supply fuses are not blown. Check that the DC voltage measurement fuses are not blown.
F07 PRESYN. UNBA.	*Pre-synchronization Unbalanced* Average output voltage imbalance of more than the value set on the screen [G7.2.1 Outp. Voltage Unbalance] for a time higher than the value set on the screen [G7.2.2 Outp.Voltage Unbalance Delay]. Output voltage IGBT unbalanced during the synchronization. Power board fuses 203-205 blown. Incorrect wiring. EMF filter damaged.	In LVT series Check Output voltage before the power transformer. Check power board fuses and replace the blown ones. Check the voltage between the fuses and neutral J208. Request technical assistance.
F08 NET.SYNC TOU	If the inverter is synchronizing with the grid and elapses a time longer than 90sec without been synchronized with the grid, the inverter will be tripped. Unbalanced values in the Display (SV 1).	Power Board has to be replaced. Request technical assistance.
	Starting Voltage is too low.	Increase Starting Voltage (G2.3).
	Bad connection at Q3.	Check screws on the AC output of the Inverter.
F09 P. SYNC. PHALO	If the inverter is synchronizing with the grid and the current of the panels exceeds 5% of rated current, the inverter will be tripped. A high DC current happened during pre-synchronization stage.	Request technical assistance.
F12 UH DESATURAT	IGBT, Gate Drive, cables or Control Board or Power Board damaged. Power semiconductor internal protection has been activated. (IGBT's)	Check the cables between Power Board and IGBT.
F13 UL DESATURAT		
F14 VH DESATURAT		Check IGBTs about visual damages.
F15 VL DESATURAT		
F16 WH DESATURAT		Request technical assistance.
F17 WL DESATURAT		
F18 D DESAT. UVW	Automatic internal protection of several of the power semiconductors has been activated. Problem in IGBT communication.	Check the resistors on the Power Board next to J209 about visual damages.
		Check and replace the fuse 800 on Power board.
	Cable on Control Board in J2004 is missing.	Install a cable to J2004 and reset the Inverter over Display.
F19 AC OVERCUR HW	Starting Voltage is too low.	Increase starting voltage in G2.3.
	Grid current has reached a dangerous level. Its value is above 150% of the inverter rated current. AC Power is stronger than DC Power.	Check DC Bus about visual damages.
		Check for the Can bus loop end resistors (HE Series)
		Request technical assistance.

DISPLAY	DESCRIPTION OR POSSIBLE CAUSE	ACTIONS
F20 PDINT	Grid and DC current have reached a dangerous level. Over current DC or AC H/W Over voltage DC H/W	Follow the instructions for F1, F3 and F19 faults. Request technical assistance.
F25 NET SEC. INV.	The grid voltages are connected with the inverted sequence.	Invert two input phases of the grid Request technical assistance.
F26 L1 LOW VOLT.	The L1 phase voltage is below the threshold set on the screen [G7.1.1 Low AC Voltage] after the time set on the screen [G7.1.2 T Low AC Voltage Delay]. Fuses F202 on the power board blown Damaged wiring.	Check de L1 voltage with an external voltmeter and the displayed voltage in the inverter. Check Fuses F200 to F202 on the power board and replace if blown. Check voltage between fuses and neutral J200. Check wiring and connectors Check the values of the network protections G7.1 Request technical assistance.
F27 L2 LOW VOLT.	The L2 phase voltage is below the threshold set on the screen [G7.1.1 Low AC Voltage] after the time set on the screen [G7.1.2 T Low AC Voltage Delay]. Fuses F201 on the power board blown Damaged wiring.	Follow the instructions F26
F28 L3 LOW VOLT.	The L3 phase voltage is below the threshold set on the screen [G7.1.1 Low AC Voltage] after the time set on the screen [G7.1.2 T Low AC Voltage Delay]. Fuses F200 on the power board blown Damaged wiring.	Follow the instructions F26
F29 L1 HIGH VOLT.	The L1 phase voltage is above the threshold set on the screen [G7.1.5 High AC Voltage] after the time set on the screen [G7.1.6 High AC Voltage Delay]. High voltage line 1	Check de L1 voltage with an external voltmeter and the displayed voltage in the inverter.(SV.1) Check voltage between fuses F200 to F202 and neutral J200. Check wiring and connectors Check the values of the electric grid protections G7.1 Request technical assistance.
F30 L2 HIGH VOLT.	The L2 phase voltage is above the threshold set on the screen [G7.1.5 High AC Voltage] after the time set on the screen [G7.1.6 High AC Voltage Delay]. High voltage line 2	Follow the instructions F29
F31 L3 HIGH VOLT.	The L3 phase voltage is above the threshold set on the screen [G7.1.5 High AC Voltage] after the time set on the screen [G7.1.6 High AC Voltage Delay]. High voltage line 3	Follow the instructions F29
F32 HIGH AC.I.L1	SW has detected that the grid phase L1 current has reached a dangerous level. Its value is above 150% of the inverter rated current. Over Current AC L1 S/W	Request technical assistance.
F33 HIGH AC.I.L2	SW has detected that the grid phase L2 current has reached a dangerous level. Its value is above 150% of the inverter rated current. Over Current AC L2 S/W	Request technical assistance.
F34 HIGH AC.I.L3	SW has detected that the grid phase L3 current has reached a dangerous level. Its value is above 150% of the inverter rated current. Over Current AC L3 S/W	Request technical assistance.
F35 LOW NET FREQ	The grid frequency is below the threshold set on the screen [G.7.1.14 Low AC Frequency] after the time set on the screen [G7.1.15 Low AC Frequency Delay]. Low grid frequency One grid phase missing Inverted phase sequence Master undefined (HE Series)	Follow the instructions F26 Follow the instructions F25 Check if a master is defined G13.4
F36 HIGH NET FRE	The grid frequency is above the threshold set on the screen [G.7.1.17 High AC Frequency] after the time set on the screen [G7.1.18 High AC Frequency Delay]. High grid frequency One grid phase missing Inverted phase sequence Master undefined (HE Series)	Follow the instructions F35
F41 NET.V.UNBALA	Grid voltage imbalance above the threshold set on the screen [G.7.1.8 AC Voltage Unbalance] after the time set on the screen [G7.1.9 AC Voltage Unbalance Delay]. Grid voltage unbalanced One grid phase missing	Follow the instructions F26
F42 NET.I.UNBALA	Grid current imbalance above the threshold set on the screen [G.7.1.11 AC Current Unbalance] after the time set on the screen [G7.1.12 AC Current Unbalance Delay].	Measure de current with an external amperemeter and compare with the displayed current in the inverter.(SV.1) Check wiring and connectors Check the values of the electric grid protections G7.1

DISPLAY	DESCRIPTION OR POSSIBLE CAUSE	ACTIONS
	Current injection unbalanced Current transformer or wiring damaged Control or power board damaged.	Replace current transformers Request technical assistance.
F44 LOW. V_L1	The L1 phase voltage is below the threshold set on the screen [G7.1.3 Low AC Vol F] after the time set on the screen [G7.1.4 TLow AC Vol]. Low voltage line 1 Fuses F202 on the power board blown Damaged wiring.	Follow the instructions F26
F45 LOW. V_L2	The L2 phase voltage is below the threshold set on the screen [G7.1.3 Low AC Vol F] after the time set on the screen [G7.1.4 TLow AC Vol]. Low voltage line 2 Fuses F201 on the power board blown Damaged wiring.	Follow the instructions F26
F46 LOW. V_L3	The L3 phase voltage is below the threshold set on the screen [G7.1.3 Low AC Vol F] after the time set on the screen [G7.1.4 TLow AC Vol]. Low voltage line 3 Fuses F200 on the power board blown Damaged wiring.	Follow the instructions F26
F50 ISLAND MODE	Loss of power of any input phase for a period exceeding 5 seconds. Power grid lost detected	Check the electrical grid conditions Request technical assistance.
F51 DSP SIZE ID	The FREESUN has not identified the model of inverter (the voltage or power of the inverter do not match) Undefined equipment Drive Select damaged	Request technical assistance.
F53 DC I LEAKAGE	The earth leakage current device has detected an anomaly.	Press 'Reset' to check, if there is still a leakage. If the module has no real Protector check for the bridge in the connection terminal Visualize the status of the SV.12 screen warning Check the cables of the photovoltaic field. Request technical assistance.
F54 DC CONTACTOR	A fault has been detected in the DC motorized contactor. Integrated switch doesn't work.	Check if the Motorized Switch is turning to ON. If it changes back to OFF immediately the integrated switched might be damaged. Request technical assistance. Visualize if appears the status of the screen warning on SV.12
	Motorized DC-Switch shows TRIPPED.	Check if there is high DC current. If the fault persists, the breaker might be damaged.
F55 AC CONTACTOR	A fault has been detected in the AC main contactor. Integrated switch doesn't work.	Visualize if appears the status of the screen warning on SV.12 Check wiring and auxiliary contactor Request technical assistance.
F56 SOFTC. CONTAC	A fault has been detected in the soft charge contactor. Integrated switch doesn't work.	Visualize if appears the status of the screen warning on SV.12 Check wiring and auxiliary connectors .Request technical assistance
F57 DC_EXTOVERV.	A fault has been detected in the DC overvoltage protection. DEHNguard at DC Input tripped.	Verify input voltage DC. Replace DEHNguard. Check wiring and auxiliary contactor Request technical assistance.
		Check DC Bus and Filter about visual damages.
F58 AC_EXTOVERV.	A fault has been detected in the AC overvoltage protection. DEHNguard at AC Input tripped.	Verify input voltage AC. Replace DEHNguard. Check wiring and auxiliary contactor Request technical assistance.
F62	Optic fiber communication fault (HE Series)	Check fiber optic bus Check fiber optic cardboards Check if a master is activated
F70 INV.SIZE ID	A fault has been detected in the inverter identification. The Size ID Card was removed or is damaged.	If card is lost, request technical assistance. Check ID Card about visual damages.
	Power board Damaged (P.S.U.) Control Board Damaged (Varnish on connector 1)	Check for Varnish on Connector 1 on the Control board, Check for the Power supply coming from the Power board.
F71 EEPROM	A fault has been detected in the static memory of the inverter. Integrated circuit fault.	Request technical assistance.
	You just updated Software and forgot to initialize all the parameters.	Initialize Parameters

DISPLAY	DESCRIPTION OR POSSIBLE CAUSE	ACTIONS
F75 AIN1 LOSS	It means that the FREESUN has stopped receiving the analogue input signal 1 being set to "YES" the screen [G6.2.8.AI1 Loss]. The inverter has lost the signal injected through that input. Cables are disconnected at X2.9-10	Check wiring about good connection and visual damages.
F76 AIN2 LOSS	It means that the FREESUN has stopped receiving the analogue input signal 2 being set to "YES" the screen [G6.2.17. AI2 Loss]. The inverter has lost the signal injected through that input. Cables are disconnected at X2.11-12	
F77 PT100 LOSS	It means that the FREESUN has stopped receiving the PT100 analog input signal being set to "YES" the screen [G6.2.21. PT100 Loss]. The inverter has lost the signal injected through that input. Cables are disconnected at X2.13-16	
F79 I.EMERG.STOP	Activation of a digital input set as an external fault has been detected. The digital input set on the screen [G6.1.1 DI1 Option] or [G6.1.2 DI2 Option] must be adjusted to 'Extern Fault (NC)'. External emergency stop	Request technical assistance.
F80 MODBUS LOSS	Trip generated by excessive delay in the serial communication. The time elapsed since the last correct frame reception has exceeded the set on the screen [G8.2.1 Time out Serial Link] Modbus Timeout reached	The Module has been inactive in the bus for the specified time. Check if the Modbus Master is sending queries periodically. Increase the value of the Modbus timeout if necessary , G8.2.1
F83 TRAFO. TEMP	The transformer temperature has reached a dangerous level. The limit of temperature of the transformer has been exceeded.	Check Temperature. If the temperature is out of specification, request technical assistance.
		Be sure that there is nothing obstructing the cooling fans (dust, papers, dirt in general) and they rotate correctly (G5.2)
		Verify that the ambient conditions are proper for the equipment.
F84 INDUCT. TEMP	The inductance temperature has reached a dangerous level. The limit of temperature of the choke has been exceeded.	Check Temperature. If the temperature is out of specification, request technical assistance.
		Be sure that there is nothing obstructing the cooling fans (dust, papers, dirt in general) and that these rotate correctly (G5.2)
		Verify that the ambient conditions are proper for the equipment.
F85 INTERN. TEMP	The internal temperature of the inverter has reached a dangerous level. The limit of internal temperature of the electronics chamber has been exceeded.	Check Temperature. If the temperature is out of specification, request technical assistance.
		Be sure that there is nothing obstructing the cooling fans (dust, papers, dirt in general) and that these rotate correctly (G5.2)
		Verify that the ambient conditions are proper for the equipment.
F86 IGBT TEMP.	The internal temperature of the power circuit (IGBTs) has reached a dangerous level above 110ºC (see screen [SV3]) Blocked or poor ventilation.	Be sure that there is nothing obstructing the cooling fans (dust, papers, dirt in general) and that these rotate correctly (G5.2)
	The limit of temperature of the IGBT has been exceeded.	Verify that the ambient conditions are proper for the equipment.
		Check Temperature. If the temperature is out of specification, request technical assistance.
F89 WATCHDOG	An unknown fault has reset the microprocessor of the control board. A fault in the microcontroller has occurred.	Disconnect and re-connect the input power of the Inverter. If the fault persists request technical assistance of Power Electronics
F92 DSP INT.FLT	Internal DSP failure	The DSP has tripon internal error. If the error persists consult with Power Electronics.

www.powerelectronics.es | www.power-electronics.com

24 Hours Technical Assistance 365 days a year +34 96 136 65 57

HEADQUARTER • VALENCIA • SPAIN	
C/ Leonardo da Vinci, 24 – 26 • Parque Tecnológico • 46980 – PATERNA • VALENCIA • ESPAÑA Tel. 902 40 20 70 • Tel. (+34) 96 136 65 57 • Fax (+34) 96 131 82 01	
BRANCHES	
CATALONIA	**BARCELONA** • Avda. de la Ferrería, 86-88 • 08110 • MONTCADA I REIXAC Tel. (+34) 96 136 65 57 • Fax (+34) 93 564 47 52
	LLEIDA • C/ Terrasa, 13 · Bajo • 25005 • LLEIDA Tel. (+34) 97 372 59 52 • Fax (+34) 97 372 59 52
CANARY ISLANDS	**LAS PALMAS** • C/ Juan de la Cierva, 4 • 35200 • TELDE Tel. (+34) 928 68 26 47 • Fax (+34) 928 68 26 47
LEVANT	**VALENCIA** • Leonardo da Vinci, 24-26 • Parque tecnológico ● 46980 ● PATERNA Tel. (+34) 96 136 65 57 • Fax (+34) 96 131 82 01
	CASTELLÓN • C/ Juan Bautista Poeta • 2º Piso · Puerta 4 • 12006 • CASTELLÓN Tel. (+34) 96 136 65 57
	MURCIA • Pol. Residencial Santa Ana • Avda. Venecia, 17 • 30319 • CARTAGENA Tel. (+34) 96 853 51 94 • Fax (+34) 96 812 66 23
NORTH	**VIZCAYA** • Parque de Actividades • Empresariales Asuarán • Edificio Asúa, 1º B • Ctra. Bilbao · Plencia • 48950 • ERANDIO • Tel. (+34) 96 136 65 57 • Fax (+34) 94 431 79 08
CENTRE	**MADRID** • Avda. Rey Juan Carlos I, 98, 4º C • 28916 • LEGANÉS Tel. (+34) 96 136 65 57 • Fax (+34) 91 687 53 84
SOUTH	**SEVILLA** • C/Arquitectura, Bloque 6 • Planta 5ª • Módulo 2 • Parque Empresarial Nuevo Torneo • 41015 • SEVILLA Tel. (+34) 95 451 57 73 • Fax (+34) 95 451 57 73
INTERNATIONAL SUBSIDIARIES	
GERMANY	**Power Electronics Deutschland GmbH** • Dieselstrasse, 77 • D·90441 • NÜRNBERG ● GERMANY Tel. (+49) 911 99 43 99 0 • Fax (+49) 911 99 43 99 8
AUSTRALIA	**Power Electronics Australia Pty Ltd** • U6, 30-34 Octal St, Yatala, • BRISBANE, QUEENSLAND 4207 • P.O. Box 6022, Yatala DC, Yatala Qld 4207 • AUSTRALIA Tel. (+61) 7 3386 1993 • Fax (+61) 7 3386 1993
BRAZIL	**Power Electronics Brazil Ltda** • Av. Imperatriz Leopoldina, 263 – conjunto 25 • CEP 09770-271 • SÃO BERNARDO DO CAMPO - SP • BRASIL • Tel. (+55) 11 5891 9612 • Tel. (+55) 11 5891 9762
CHILE	**Power Electronics Chile Ltda** • Los Productores # 4439 – Huechuraba • SANTIAGO • CHILE Tel. (+56) (2) 244 0308 · 0327 · 0335 • Fax (+56) (2) 244 0395 Oficina Petronila # 246, Casa 19 • ANTOFAGASTA • CHILE Tel. (+56) (55) 793 965
CHINA	**Power Electronics Beijing** • Room 606, Yiheng Building • No 28 East Road, Beisanhuan • 100013, Chaoyang District • BEIJING • R.P. CHINA Tel. (+86 10) 6437 9197 • Fax (+86 10) 6437 9181 **Power Electronics Asia Ltd** • 20/F Winbase Centre • 208 Queen's Road Central • HONG KONG • R.P. CHINA
KOREA	**Power Electronics Asia HQ Co** • Room #305, SK Hub Primo Building • 953-1, Dokok-dong, Gangnam-gu • 135-270 • SEOUL • KOREA Tel. (+82) 2 3462 4656 • Fax (+82) 2 3462 4657
INDIA	**Power Electronics India** • No 25/4, Palaami Center, • New Natham Road (Near Ramakrishna Mutt),• 625014 • MADURAI Tel. (+91) 452 452 2125• Fax (+91) 452 452 2125
ITALY	**Power Electronics Italia Srl** • Piazzale Cadorna, 6 • 20123 • MILANO • ITALIA Tel. (+39) 347 39 74 792
MEXICO	**P.E. Internacional Mexico S de RL** • Calle Cerrada de José Vasconcelos, No 9 • Colonia Tlalnepantla Centro • Tlalnepantla de Baz • CP 54000 • ESTADO DE MEXICO Tel. (+52) 55 5390 8818 • Tel. (+52) 55 5390 8363 • Tel. (+52) 55 5390 8195
NEW ZEALAND	**Power Electronics New Zealand Ltd** • 12A Opawa Road, Waltham • CHRISTCHURCH 8023 • P.O. Box 1269 CHRISTCHURCH 8140 Tel. (+64 3) 379 98 26 • Fax.(+64 3) 379 98 27
UNITED KINGDOM	**Power Electronics UK Pty Ltd**• Wells House, 80 Upper Street, Islington, • London, N1 0NU • 147080 Islington 5 Tel. (+34) 96 136 65 57 • Fax (+34) 96 131 82 01
SOUTH AFRICA	**Power Electronics South Africa Pty Ltd** • Central Office Park Unit 5 • 257 Jean Avenue • Centurion 0157 Tel. (+34) 96 136 65 57 • Fax (+34) 96 131 82 01

POWER ELECTRONICS®

www.power-electronics.com

Appendix B

Appendix B

The intention of this project was not to provide a useable program that but instead to explore the data provided and document well the code required to do so along with model iteration. In accordance with this, Jupyter notebooks can be found at the following Git Repositories

- https://github.com/jakec338/Solar_Predictive_Maintenance/blob/master/HydesPredictiveMaintenance_R_notebook_copy.ipynb
- https://github.com/jakec338/Solar_Predictive_Maintenance/blob/master/Hydes_Predictive_Maintenance_Models.ipynb

Bibliography

[1] *Energy statistics - an overview*.
http://web.archive.org/web/20080207010024/https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Energy_statistics_-_an_overview.

[2] *Python vs r – who is really ahead in data science, machine learning?*
http://web.archive.org/web/20080207010024/https://www.kdnuggets.com/2017/09/python-vs-r-data-science-machine-learning.html, journal=KDnuggets, year=2017, note = Accessed: 03/07/2018.

[3] *Data analytics for solar plant performance improvement*.
http://web.archive.org/web/20080207010024/http://algoengines.com/wp-content/uploads/2016/01/Data-Analytics-for-Solar-Farm-Performance-Improvement.pdf, 2017.
Accessed: 20/07/2018.

[4] *Energy statistics - an overview*.
http://web.archive.org/web/20080207010024/https://businessanalystlearnings.com/ba-techniques/2013/3/5/moscow-technique-requirements-prioritization, 2017.
Accessed: 15/07/2018.

[5] *A peek into xgboost with python*.
http://web.archive.org/web/20080207010024/https://www.channels.elastacloud.com/channels/championing-data-science/a-peek-into-xgboost-with-python, 2017.
Accessed: 03/08/2018.

[6] *Predictive analytics for solar operators*.
http://web.archive.org/web/20080207010024/https://egen.solutions/assets/Predictive_Analytics_for_Solar_Operators.pdf, 2017.
Accessed: 19/07/2018.

[7] *Quintas energy*.
http://web.archive.org/web/20080207010024/https://www.quintas.com, 2017.

Accessed: 18/07/2018.

[8] *Solar panel warranties*. `http://web.archive.org/web/20080207010024/https://www.renewableenergyhub.co.uk/solar-panels/solar-panel-warranty-insurance-maintenance.html`, 2017. Accessed: 18/07/2018.

[9] D. BAILEY AND E. WRIGHT, *Practical SCADA for industry*, Elsevier, 2003.

[10] B. K. BARADWAJ AND S. PAL, *Mining educational data to analyze students' performance*, arXiv preprint arXiv:1201.3417, (2012).

[11] P. BEITER, M. ELCHINGER, AND T. TIAN, *2016 renewable energy data book*, tech. rep., NREL, 2017.

[12] C. M. BISHOP, *Machine learning and pattern recognition*, 2006.

[13] C. BÖHRINGER, F. LANDIS, T. REAÑOS, AND M. ANGEL, *Economic impacts of renewable energy promotion in germany.*, Energy Journal, 38 (2017).

[14] L. BREIMAN, *Random forests*, Machine learning, 45 (2001), pp. 5–32.

[15] M. M. BREUNIG, H.-P. KRIEGEL, R. T. NG, AND J. SANDER, *Lof: identifying density-based local outliers*, in ACM sigmod record, vol. 29, ACM, 2000, pp. 93–104.

[16] S. V. BUUREN AND K. GROOTHUIS-OUDSHOORN, *mice: Multivariate imputation by chained equations in r*, Journal of statistical software, (2010), pp. 1–68.

[17] N. V. CHAWLA, K. W. BOWYER, L. O. HALL, AND W. P. KEGELMEYER, *Smote: synthetic minority over-sampling technique*, Journal of artificial intelligence research, 16 (2002), pp. 321–357.

[18] T. CHEN AND C. GUESTRIN, *Xgboost: A scalable tree boosting system*, in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, ACM, 2016, pp. 785–794.

[19] K. L. CHOPRA, P. D. PAULSON, AND V. DUTTA, *Thin-film solar cells: an overview*, Progress in Photovoltaics: Research and Applications, 12 (2004), pp. 69–92.

[20] P. DAUGHERTY, P. BANERJEE, W. NEGM, AND A. E. ALTER, *Driving unconventional growth through the industrial internet of things*, accenture technology, (2015).

[21] E. . I. S. DEPARTMENT FOR BUSINESS, *National Statistics Solar PV deployment: January 2018*, 2018.

[22] C.-D. Dumitru and A. Gligor, *Scada based software for renewable energy management system*, Procedia Economics and Finance, 3 (2012), pp. 262–267.

[23] N. Henke, J. Bughin, M. Chui, J. Manyika, T. Saleh, B. Wiseman, and G. Sethupathy, *The age of analytics: Competing in a data-driven world*, McKinsey Global Institute, 4 (2016).

[24] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, *Long short term memory networks for anomaly detection in time series*, in Proceedings, Presses universitaires de Louvain, 2015, p. 89.

[25] E. Nugent, *Why solar scada is taking center stage in grid modernization*, www.renewableenergyworld.com, (2017).

[26] J. Pastor-Pellicer, F. Zamora-Martínez, S. España-Boquera, and M. J. Castro-Bleda, *F-measure as the error function to train neural networks*, in International Work-Conference on Artificial Neural Networks, Springer, 2013, pp. 376–384.

[27] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt, *Support vector method for novelty detection*, in Advances in neural information processing systems, 2000, pp. 582–588.

[28] P. Senin, *Dynamic time warping algorithm review*, Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA, 855 (2008), pp. 1–23.

[29] D. Shiffman, *The Nature of Code: Simulating Natural Systems with Processing*, Daniel Shiffman, 2012.

[30] A. Shmilovici, *Support Vector Machines*, Springer US, Boston, MA, 2010.

[31] X.-S. Si, W. Wang, C.-H. Hu, and D.-H. Zhou, *Remaining useful life estimation –a review on the statistical data driven approaches*, European Journal of Operational Research, 213 (2011), pp. 1–14.

[32] G. A. Susto, A. Schirru, S. Pampuri, S. McLoone, and A. Beghi, *Machine learning for predictive maintenance: A multiple classifier approach*, IEEE Transactions on Industrial Informatics, 11 (2015), pp. 812–820.

[33] R. Vlasveld, *Introduction to one-class support vector machines*, City, (2013).

[34] R. Wirth and J. Hipp, *Crisp-dm: Towards a standard process model for data mining*, in Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, Citeseer, 2000, pp. 29–39.

[35] D. M. Witten and R. Tibshirani, *Survival analysis with high-dimensional covariates*, Statistical methods in medical research, 19 (2010), pp. 29–51.

[36] A. Zaher, S. McArthur, D. Infield, and Y. Patel, *Online wind turbine fault detection through automated scada data analysis*, Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology, 12 (2009), pp. 574–593.