Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/31/20230142

Comparison of machine learning models for the estimation of solar power generation in Australia

Zhuowen Han

Sydney Smart Technology College, Northeastern University, Qinhuangdao, Hebei, 066000, China.

hanzhuowen@hbuas.edu.cn

Abstract. Solar power is a renewable energy source that contributes a lot to improving the environmental sustainability of energy production. In Australia, rooftop solar panel systems are installed for power generation. By seeking the optimal installation strategy, people could develop more efficient technologies and systems. In this work, data related to power generation was gathered between January 1 and December 31, 2022. This article only analyzes qualitative data in the dataset. Firstly, the support vector machine (SVM) model is used to compare the results obtained by removing only one qualitative feature at a time with the results obtained by including all features. It is concluded that the shading condition feature has a significant impact on power generation, and then. By comparing the accuracy of K-nearest neighbor (KNN), Logistic Regression, SVM with Radial Basis Function (RBF) kernel, and Decision Tree, it was found that SVM with RBF kernel is the most suitable classification method for the feature of the shading condition.

Keywords: solar power generation, machine learning, support vector machine.

1. Introduction

In order to transform light energy directly into electrical energy, rooftop solar panel systems are installed in Australia. For each system, there are three parts: solar panels, controllers, and inverters [1]. These systems could produce stable and ample solar energies, once these solar cells are installed and arranged in series. By connecting theses power controllers and other corresponding parts, a photovoltaic power generation device could be produced.

Both from a global and a Chinese perspective, conventional energy is extremely scarce. A significant role in the long-term energy strategy is played by solar energy, which is an unrenewable energy source that is infinitely pure, completely safe, relatively extensive, truly long-lasting, and maintenance-free for humans [2, 3]. The advantages of photovoltaic power generation over commonly used thermal power generation systems are primarily reflected in: (1) no risk of depletion; (2) safety and reliability; zero noise; zero pollution discharge; and absolute cleanliness (no pollution). Therefore, this effort has chosen to investigate this cutting-edge green technology.

DOI: 10.54254/2755-2721/31/20230142

2. Method

2.1. Dataset

The information used in this study comes from an Australian website where visitors can enter details about their home. Moreover, the estimated power generation will also be delivered [4]. In this dataset, the information is collected between Jan 1st to Dec 31st in 2022 from the buildings where solar panels are already installed. The purpose of this work is to investigate the effects of each feature on the final generation using the concept of control variates. The strategy is to compare the acquired results to those including all features while simply using the support vector machine classification method, removing one feature at a time.

2.2. Models

2.2.1. Support vector machine (SVM). It is widely used to address binary classification issues. The goal is to translate a vector into a higher-dimensional space with a defined maximum spacing hyperplane, with two parallel decision hyperplanes. The distance between the two parallel hyperplanes is increased when they are separated. Assume that the total error of the classifier is decreased in direct proportion to the distance or difference between parallel Hyperplanes [5,6]. It has a number of benefits. In the beginning, it makes use of kernel functions that can map to high-dimensional spaces. Second, it employs kernel functions to address issues with nonlinear classification. Thirdly, the classification principle is straightforward: the sample should be kept as far away from the decision surface as possible. Fourthly, it has a good effect on classification. It does, however, have drawbacks. First of all, implementing the SVM algorithm for extensive training data is challenging. Second, using SVM to address numerous classification issues is challenging. Thirdly, it is sensitive to the choice of parameters and kernel functions as well as missing data.

2.2.2. K nearest neighbors (KNN). Based on the separation between various features, KNN is categorized. The sample will also fall into this group if most of the K adjacent samples in the feature space, where K is often an integer not exceeding 20, do the same. Various K could lead to significant impact on the KNN algorithm's results. When the training set's data and labels are known, enter the test data, compare its features to those in the training set, and pick the first K data that most closely resembles the training [7,8]. The K data category with the greatest number of occurrences corresponds to the test data category.

2.2.3. Decision tree. The choice of categorisation A tree structure known as a "tree model" specifies how instances are categorized. A decision tree has directed edges and nodes with two types, including internal nodes and leaf nodes. Leaf nodes indicate a class, whereas internal nodes represent a feature or characteristic. A prediction analysis model described as a tree structure, such as a binary tree or a multi-tree, is referred to as a decision tree. Sort instances into categories by putting them in a root-to-leaf order [9,10]. The classification to which the instance belongs is shown by the leaf node. Each node in the tree represents the testing of a specific instance attribute, and each branch that follows the node represents a potential value of that attribute.

2.3. Evaluation indexes

The percentage of accurately identified samples over all samples is known as accuracy. A statistical metric for all samples is accuracy. The easiest and most logical evaluation indication in classification problems is accuracy, although it has apparent drawbacks. For instance, if the percentage of negative samples is 99%, the classifier can still achieve 99% accuracy by classifying all samples as negative samples. Therefore, the categories with a high proportion frequently become the most crucial element impacting accuracy when the proportion of samples in different categories is very uneven.

DOI: 10.54254/2755-2721/31/20230142

Precision is the ratio of accurately identified positive samples to the total number of samples the classifier identified as positive samples. With an emphasis on the statistics of data that the classifier determines to be positive, accuracy is a statistical measure of some samples.

The percentage of accurately identified positive samples to the actual number of positive samples is referred to as the recall rate. Another statistical metric for some samples that concentrates on the statistics of truly positive samples is recall rate.

3. Result and discussion

Before starting to build the model, the dataset is separated into a training set and a testing set in an 8:2 ratio to prevent data snooping bias.

3.1. Result of feature importance analysis

Firstly, when the SVM model is selected to classify a dataset containing all features, the results obtained are demonstrated in Figure 1.

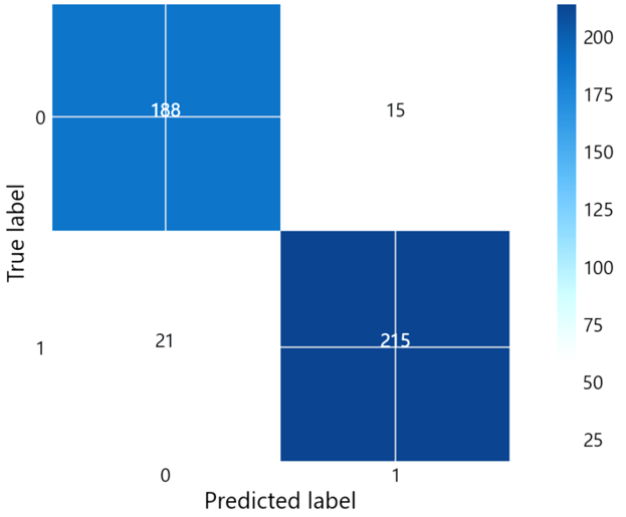

Figure 1. Confusion matrix of SVM (Figure credit: Original).

Through the above Confusion matrix, results show that when all features are included, the accuracy of using SVM for classification is very high. After further research, it is found that all evaluation indicators of the model have achieved good results.

Confusion matrix in Figure 2 demonstrates the result without shading, it could be observed that the number of prediction errors has increased significantly, and here it could be roughly estimated that the accuracy of the model and other evaluation indicators will decline. By observing the Confusion matrix in the right part of Figure 2, it could be found that the accuracy of the model is still high, but this is only a rough estimate. From the following table, results demonstrate that this is no different from the results containing all features, and it could be temporarily concluded that this feature of financial will not have a great impact on the final power generation.

DOI: 10.54254/2755-2721/31/20230142

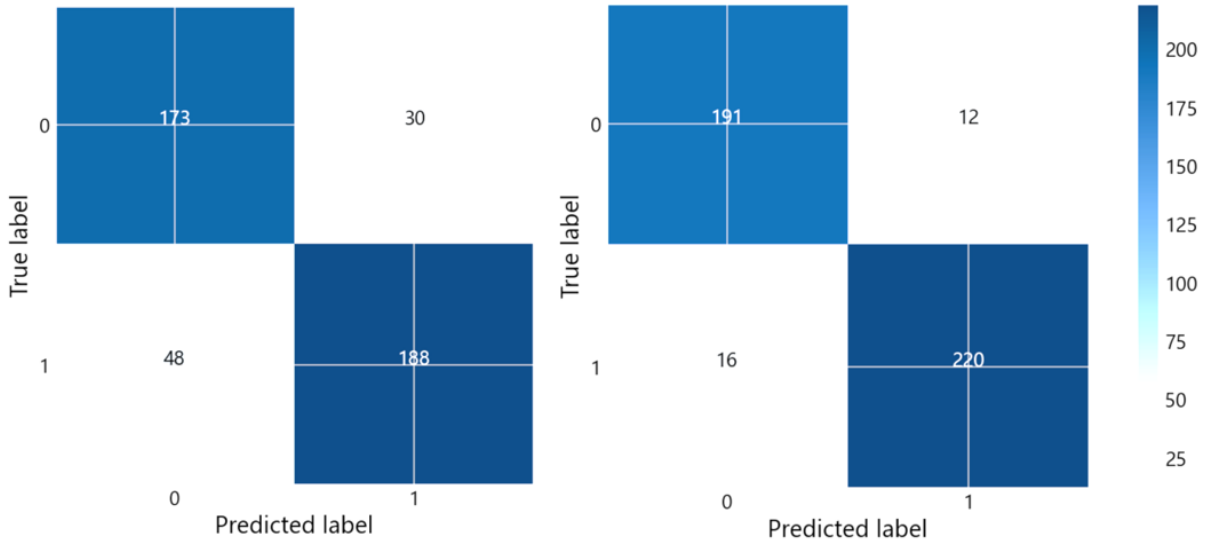

Figure 2. Confusion matrix without shading (left) and without financial (Figure credit: Original).

Table 1 compares the result of the original SVM together with the result without shading and financial information. From the table below, it could be observed that the accuracy, regression rate, and f1 score of the support vector machine model are all above 90%. This indicates that the model can effectively classify datasets.

Table 1. Result comparison of using full features, feature without shading, and feature without financial.

	Accuracy	Precision	Recall	F1-score
Original	0.92	0.92	0.92	0.92
Without shading	0.82	0.82	0.82	0.82
Without financial	0.94	0.94	0.94	0.94

As expected, compared to the results obtained by including all features, all evaluation indicators showed varying degrees of decline, with accuracy dropping from 92% to 78%, and recall rates dropping from 93% and 91% to 85% and 80%. These changes all reflect a significant decrease in the accuracy of the model after removing the shading feature, which also reflects the important role of the shading feature in the final power generation. Now it is known that among these features, shading has the greatest impact on power generation, so this work wants to continue studying whether the results will change if the parameters of the SVM model are changed and find the optimal result.

Shading condition indicate the area of solar panels covered by adjacent trees of other buildings. There are 3 classes in this feature, significant, partial and none. First a pair plot of shading condition is demonstrated.

Although there are 3 classes in shading condition, it can be easily observed that this feature is feasible to be classified. The boundaries between each 2 classes are relatively clear. Digital features, latitude, roof pitch, roof azimuth, year, panel capacity and generation, are used to classify the shading condition. In the following parts, these digital features are also used to classify other factors.

Four methods are used to classify shading condition, which are SVM with different kernels, logistic regression, decision tree, and KNN. The optimal hyperparameter for each algorithm is explored in this report. The cleaned dataset is divided at ratio of 8:2 for training and testing. The training set and testing set are chosen randomly. This treatment is applied on other factors for classification.

DOI: 10.54254/2755-2721/31/20230142

3.2. Result of decision tree

There are 2 criteria for decision tree in scikit learn, which are gini and entropy. And choosing to consider weight in decision tree or not is also an interesting topic. Choosing weight means that the sample sizes of different classes are taken into consideration. This report makes a comparison between different choices of decision tree. Here Table 2 presents the accuracy and recall scores of different decision trees.

Table 2. Result of decision tree.

	Gini	Gini, weight	Entropy	Entropy, weight
Accuracy	0.8617	0.8447	0.8738	0.8471
Recall	0.8102	0.7432	0.7984	0.7391

It is shown that decision tree taking entropy without considering weight achieves the highest accuracy score, but Gini decision tree without considering weight has the best overall result. So in the following discussion, Gini decision tree without considering weight is chosen.

3.3. Result of KNN

The choice of the number of neighborhood points is crucial in KNN algorithm. To find the optimal number of neighborhood points, Figure 3 demonstrates the values of accuracy scores and recall scores change with the number of neighborhood points is presented. The range of number of neighborhood points is from 1 to 50.

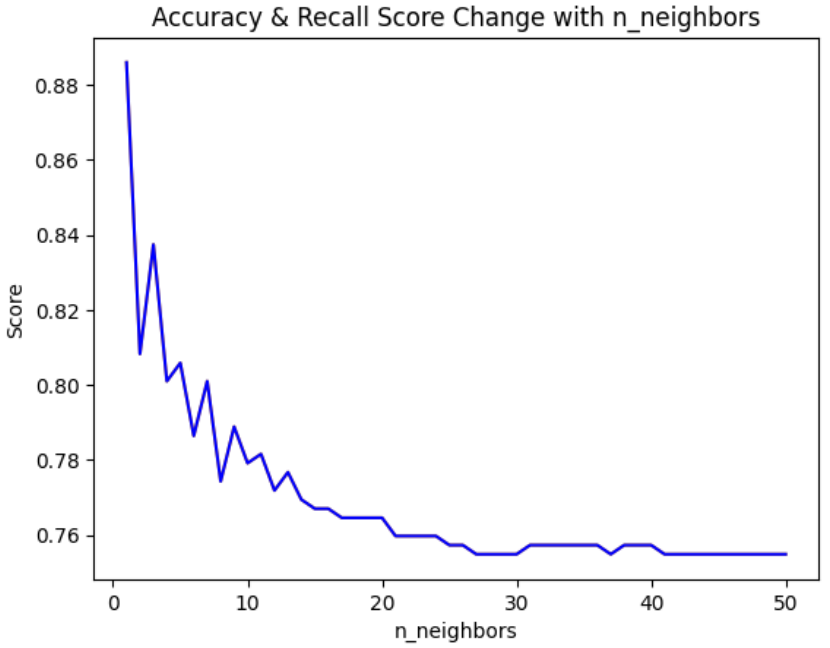

Figure 3. Result of KNN with different number of neighbors (Figure credit: Original).

It's clearly shown in the graph that when the number of neighborhood point is 1, the accuracy and recall scores are highest. In the following discussion, nearest neighbor method is selected.

3.4. Result of SVM

The choice of hyperparameter, penalty coefficient C, and the choice of kernel are both very important in SVM method. First, when the penalty coefficient is 5, SVM with different kernels are tested on the dataset. The kernels to be selected are polynomial, sigmoid, RBF and linear kernel.

DOI: 10.54254/2755-2721/31/20230142

In addition, the impact of standardization on accuracy and recall scores is also tested on different SVM methods. The accuracy and recall scores are listed in Table 3.

Table 3. Result of SVM with different settings.

		Poly	Sigmoid	RBF	Linear
No	Accuracy	0.7646	0.6699	0.7816	--
No	Recall	0.5139	0.3872	0.7516	--
Yes	Accuracy	0.9515	0.7354	0.9976	0.8617
Yes	Recall	0.9756	0.6279	0.9990	0.8546

The SVM with RBF kernel achieves the best result. And it's also clear that standardization can improve both the accuracy and recall scores of 4 SVM methods significantly. Especially for linear kernel SVM, the result is even unachievable without standardization.

Penalty coefficient is the hyperparameter that represents the generalization capability of the SVM model. When the penalty efficient C is smaller, the generalization capability of the model is better. So, it is necessary to find out the smallest C achieving the best overall result. To test the value changes of accuracy and recall scores with the change of penalty coefficient, the selective values of penalty coefficient are 0.5, 0.6, 0.7, 0.8, 0.9 and 1 to 7. Figure 4 demonstrates the changes are presented. SVM with RBF kernel is chosen to be the testing model.

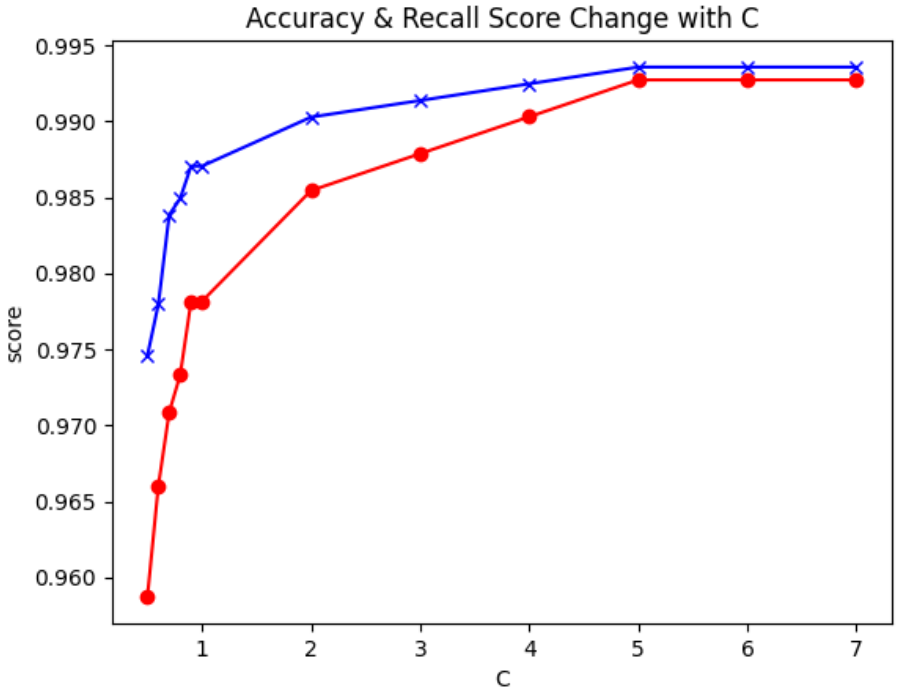

Figure 4. Result of different penalty score of SVM (Figure credit: Original).

Therefore, when the penalty coefficient is about 5, the overall result is the best. In the following discussion, SVM with RBF kernel and C=5 is selected to be compared with other optimal models.

3.5. Result comparison

The nearest neighbor method, logistic regression, RBF kernel SVM with the penalty coefficient valuing 5 and the gini decision tree without weight are compared on the classification of shading condition. The results are demonstrated in Table 4.

DOI: 10.54254/2755-2721/31/20230142

Table 4. Result comparison of various models.

		KNN(1)	Logistic Regression	SVM(RBF)	Decision Tree
No	Accuracy	0.7646	0.6699	0.7816	--
No	Recall	0.5139	0.3872	0.7516	--
Yes	Accuracy	0.9515	0.7354	0.9976	0.8617
Yes	Recall	0.9756	0.6279	0.9990	0.8546

Obviously, the SVM method has the highest accuracy score and recall score. Additionally, standardization can improve the accuracy and recall scores of all the models except for decision tree. The confusion matrix of RBF SVM is also provided.

4. Conclusion

This work aims at estimating the solar power generation conditions in Australia. In the first stage, this work newly used SVM method to classify the data, and then compared various qualitative features with a method similar to the Control variates to get the shading feature, which has a great impact on the final power generation, so the roof without shadow is more suitable for installing solar panels than other roofs. In the second section, the nearest neighbor method, logistic regression, RBF kernel SVM with the penalty coefficient valuing 5 and the gini decision tree without weight are compared on the classification of shading condition. It is evident that the SVM method has the highest accuracy, so the SVM method is selected for classification. In the future, more advanced models such as deep learning-based models could be validated on this dataset for exploring the access to better performance.

References

[1] Hosenuzzaman, M., Rahim, N. A., Selvaraj, J., Hasanuzzaman, M., Malek, A. A., & Nahar, A. (2015). Global prospects, progress, policies, and environmental impact of solar photovoltaic power generation. Renewable and sustainable energy reviews, 41, 284-297.

[2] Kannan, N., & Vakeesan, D. (2016). Solar energy for future world:-A review. Renewable and sustainable energy reviews, 62, 1092-1105.

[3] Gong, J., Li, C., & Wasielewski, M. R. (2019). Advances in solar energy conversion. Chemical Society Reviews, 48(7), 1862-1864.

[4] Solar power generation forecast. URL: https://www.kaggle.com/code/pythonafroz/solar-power-generation-forecast-with-99-auc/notebook. Last accessed 2023/07/12.

[5] Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In Machine learning, 101-121.

[6] Tanveer, M., Rajani, T., Rastogi, R., Shao, Y. H., & Ganaie, M. A. (2022). Comprehensive review on twin support vector machines. Annals of Operations Research, 1-46.

[7] Peterson, L. E. (2009). K-nearest neighbor. Scholarpedia, 4(2), 1883.

[8] Jiang, L., Cai, Z., Wang, D., & Jiang, S. (2007). Survey of improving k-nearest-neighbor for classification. In Fourth international conference on fuzzy systems and knowledge discovery (FSKD 2007) 1, 679-683.

[9] Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2(01), 20-28.

[10] Zhou, H., Zhang, J., Zhou, Y., Guo, X., & Ma, Y. (2021). A feature selection algorithm of decision tree based on feature weight. Expert Systems with Applications, 164, 113842.