Diversity Factor Prediction for Distribution Feeders with Interpretable Machine Learning Algorithms

Wenyu Wang, Nanpeng Yu, Jie Shi
Department of Electrical and Computer Engineering
University of California, Riverside
Riverside, USA

Nery Navarro
Grid Modernization
Southern California Edison
Pomona, USA

***Abstract*—The maximum diversified demand is an important factor to consider when utilities design new distribution systems. To estimate the maximum diversified demand, engineers need to make an estimate of the diversity factor (DF). In practice, electricity utility companies usually estimate the DF using DF tables, in which the DF changes with the number of customers. However, besides the number of customers, DF also depends on many other factors, such as customer type, weather, demographics, and socioeconomic conditions. Ignoring these factors, DF tables have limited accuracy. In addition, engineers cannot interpret or understand how various factors affect the DF. In this paper, by leveraging supervised machine learning algorithms, we build comprehensive DF prediction models that take a variety of factors into account. These models show high prediction accuracy and interpretabilty when applied to real-world distribution feeders. Using the interpretation method called SHapley Additive exPlanations, we quantify the importance of different features in determining DFs. Finally, we offer more insights into how various factors affect DFs.**

***Index Terms*—Distribution circuit planning, diversity factor, interpretable machine learning, SHapley Additive exPlanations, supervised machine learning.**

I. Introduction

The maximum diversified demand, i.e., the maximum of the sum of demands of a group of electricity customers over a particular period, is one of the most important factors to consider when utilities develop plans to build new distribution systems. It is very important to the design of both network topology and the ratings of equipment. Underestimating the maximum diversified demand will cause reliability and safety issues. If the peak load exceeds the circuit rating, then equipment such as transformers and cables will be overloaded, which results in shortened lifespan and premature failure. Overestimating the maximum diversified demand often leads to installation of oversized distribution system equipment and under-utilization of system assets.

The maximum diversified demand is usually estimated by using the maximum noncoincident demand and the diversity factor (DF), which is defined as follows [1]:

$$\textit{Diversity factor} = \frac{\textit{Maximum noncoincident demand}}{\textit{Maximum diversified demand}}. \quad (1)$$

Here, the maximum noncoincident demand is the sum of each individual customer's maximum demand. Obviously, DF is greater than or equal to 1. A higher DF means that customers have more diversified usage patterns and their individual maximum loads have less coincidence in time. In general, as the number of customers increases, DF first increases and then gradually levels off.

The maximum noncoincident demand is straightforward to estimate because an individual customer's maximum demand is the customer's electric service rating, which can be obtained by survey. Thus, the key problem is how to estimate DF.

In practice, DF is often estimated based on a simple relationship. Engineers estimate DF by referring to a DF table, in which the DF value varies with the number of customers. The DF table is often derived by utilities through load surveys from a few groups of customers in the distribution system [1]. In the load survey, the maximum demand of each individual customer and their maximum diversified demand are recorded. However, DF is influenced by many other factors, such as customer demographics and climate conditions. Thus, DF tables, which ignore these factors, have limited accuracy. Furthermore, engineers cannot interpret or explain how various factors affect the DF.

Researchers have studied different aspects of DF and demand diversity. However, very few research efforts have focused on developing comprehensive and interpretable prediction models for DF which account for various input features [2]. Early research [3] models DF as a function of the number of customers. Different DF functions are derived based on time of the year, day of the week, and whether electric-heating is used. Ref. [4] studies the distribution of DF and shows that DF follows gamma distributions rather than Gaussian distributions. Ref. [5] studies a metric called after-diversity maximum demand of n customers (ADMD^n), which is closely related to DF and demand diversity. This work shows that ADMD^n is affected by customers' household occupancy and wealth levels. In [6], a variable truncated R-vine copulas method is used to estimate the maximum diversified demand of customers of different household occupancy and wealth levels.

In this paper, we develop comprehensive models based on supervised machine learning algorithms to predict the DF of distribution feeders, accounting for a variety of influential factors, such as customer type, weather, demographics, and socioeconomic conditions. The machine learning algorithms not only yield high prediction accuracy on real-world distribution feeders but also provide useful insights on how input

features influence DF. Using the interpretation method called SHapley Additive exPlanations (SHAP) [7], we identify the key factors that affect the DF.

The rest of the paper is organized as follows: Section II explains the machine learning methodologies used to develop and interpret the DF prediction model. Section III summarizes the real-world distribution feeders and influential factors used to construct the dataset for the DF prediction model. Section IV shows the DF prediction performance and provides interpretation for the model. Section V states the conclusions.

II. Machine Learning Methodologies for DF Prediction Models

We adopt supervised machine learning algorithms to build the DF prediction model, which maps the input features to the output (i.e., DF of a feeder). In supervised machine learning, a model learns its mapping from a training dataset, which are samples of correct input-output pairs. Mean square error (MSE) is used to measure the model prediction performance. The details of DF prediction model development are provided in Section II-A. To interpret the prediction model, we use a method called SHAP [7] to identify the most important input features that influence the DF prediction. The details of SHAP are explained in Section II-B

A. Supervised Machine Learning Algorithms

To estimate DF of distribution feeders, we adopt 3 types of supervised machine learning algorithms: feed forward neural network (FNN), gradient boosted trees (GBT), and random forest. We choose these 3 algorithms, because they are widely used in the machine learning field and achieve great results in various problems. We further improve FNN by adding dropout layer(s) and introducing network pruning. Thus, in total, we deploy 6 algorithms: FNN, FNN+dropout, FNN+pruning, FNN+dropout+pruning, GBT, and random forest. The overall framework for building and evaluating DF prediction models consists of 3 steps. First, we preprocess the dataset and split it into training, validation, and test datasets. Second, for each of the 6 models, we train the model and tune the model's hyperparameters. Third, we evaluate the performance of the 6 prediction models using the test dataset. Due to the underlying randomness in the training and model initialization processes, we train each model 10 times and report the average model prediction errors. The technical details related to the supervised machine learning algorithms are presented below.

1) FNN: Our base FNN consists of three components: an input layer of 45 nodes, three hidden layers of 200 nodes, and an output layer of 1 node. Each node has directed connections to the nodes of the subsequent layer and each connection has a corresponding weight. In the input layer, each node corresponds to an input variable. In the hidden layer, each node takes in the weighted sum of nodes from the previous layer (plus a bias term) and produces an output value by the ReLU activation function. The output layer is a linear function of the nodes in the last hidden layer. When training FNN and its variants, we use early stopping with patience=200 epochs.

2) Network Pruning: Pruning removes unnecessary branches to improve the performance of FNN. We adopt an innovative pruning method called lottery ticket [8], [9]. This pruning method comprises the following steps: a) randomly initialize a neural network with weights w_i; b) train the neural network, reaching the trained weights w_f; c) prune $p\%$ of the weights that have the smallest w_f in magnitude, i.e., set the pruned weights to 0; d) reset the unpruned weights in w_f to their initial values in w_i (i.e., winning tickets) and retrain the network while keeping the pruned weights to 0. It is believed that pruning produces a sparse neural network with less connections, which can reduce overfitting. In addition, the winning tickets may discover a good initialization point that already lies in the randomly initialized network.

3) GBT: GBT is an ensemble learning method, which consists of a series of decision trees. The summed/aggregated prediction of the decision trees are used as the output. The GBT is trained by adding one tree a time while keeping the existing trees unchanged. Each new tree is trained using a gradient descent procedure so that the loss of the ensemble model is reduced. To avoid overfitting in the training process, we use early stopping technique with patience=200 to decide when to stop adding trees.

4) Random Forest: Random forest is another widely used ensemble learning method. It outputs the average prediction of multiple decision trees, which are fitted to various subsets of the dataset. Different from GBT, which trains a new tree based on the existing ones, random forest trains trees that are almost independent.

5) Data Preprocessing and Split: Every numerical input feature is standarized, i.e., centered and normalized by its standard deviation. This standarization shifts and rescales feature variables to similar ranges and thus improves convergence in the training process. Every categorical feature variable is represented by one-hot encoding. In our problem, only the climate zone featue is a categorical variable. For input features that are linearly dependent, we remove one of them. For example, the ratios of population in different age ranges sum up to 1. Thus, we remove one of the ratios. Such features, called redundant features, are highly correlated with other features, so they do not provide relevant information. It is a common practice to remove them in machine learning.

64% of the samples in the dataset are used to train the prediction models. 16% of the samples are used as the validation dataset for hyperparameter tuning and early stopping. The remaining 20% of the samples are used as the test set to evaluate the models' prediction performance.

6) Hyperparameter Tuning: Hyperparameters are the settings and parameters that control the configuration and influence the performance of machine learning algorithms. Following the common pratice, we use the validation dataset to tune the hyperparameters. Under different hyperparameter settings, each model is trained 10 times using the training dataset and then evaluated on the validation set. For each of the 6 prediction models, the hyperparameter setting with the lowest average validation MSE is selected.

The possible hyperparameter settings for all 6 models are listed in Table I. Every combination of the hyperparameter settings is examined when tuning the hyperparameters. For the model FNN+Dropout, all setting combinations between FNN and dropout are examined. For the FNN+Pruning model and FNN+Dropout+Pruning, we fix the hyperparameter settings already tuned for FNN and FNN+Dropout, and only tune the network pruning rate $p\%$.

TABLE I: Summary of Hyperparameters and Their Settings

Model	Hyperparameters and Their Possible Configurations
FNN	Batch size = [5,10,50,100]; optimizer = [Stochastic Gradient Descent, RMSprop, Adagrad, Adadelta, Adam, Adamax, Nadam].
Dropout	Input layer dropout ratio = [0.05, 0.1, 0.15, 0.2, 0.3]; hidden layer dropout ratio = [0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5].
Pruning	Pruning rate $p\%$ = [25%, 50%, 75%, 85%, 95%, 98%, 99%].
GBT	Max. tree depth= [2,4,6,8]; learning rate = [0.01, 0.02, 0.05, 0.1]; subsample ratio of training instances = [0.3, 0.4, ..., 0.8]; subsample ratio of features for each split = [0.1, 0.2, ..., 1].
Random Forest	Number of trees= [5,10,50,100,500,1000,5000,10000,100000]; max. number of features to consider for the best split = [m, $\log_2(m)$, $\sqrt{m}$, $m/3$] (m: the total number of features).

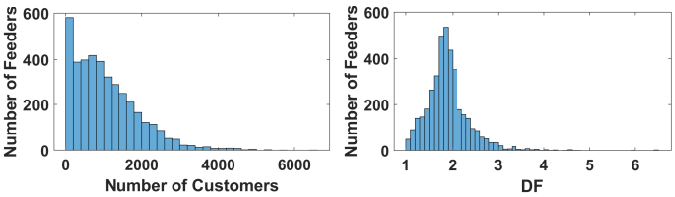
B. The SHAP Method for Model Interpretation

It is important to understand how DF of a distribution feeder is influenced by different features. In this paper, we use the SHAP [7] framework to interpret the DF prediction models. The SHAP framework has a solid theoretical foundation in cooperative game theory. It calculates each input feature's contribution to the model's output so that the influence on the output can be fairly distributed to the input features. The SHAP framework is model-agnostic, meaning that it does not require the knowledge of the model structure. Thus, SHAP works well with all types of prediction models.

The inner workings of SHAP can be explained as follows. Suppose we have a prediction model $y = f(\boldsymbol{x})$, where $\boldsymbol{x} = [x_1, ..., x_m]$ is the input feature vector and y is the model output. All the samples of $\boldsymbol{x}$ form a set $\boldsymbol{X}$. For any sample $\boldsymbol{x}^{(i)} = [x_1^{(i)}, ..., x_m^{(i)}] \in \boldsymbol{X}$, SHAP calculates a vector $\boldsymbol{\phi}^{(i)} = [\phi_1^{(i)}, ..., \phi_m^{(i)}]$ representing the contribution of each input feature in $\boldsymbol{x}^{(i)}$, such that $\sum_{j=1}^{m} \phi_j^{(i)} = f(\boldsymbol{x}^{(i)}) - E_{\boldsymbol{x2X}}(f(\boldsymbol{x}))$. Here, $E_{\boldsymbol{x2X}}(f(\boldsymbol{x}))$ is the expectation of $f(\boldsymbol{x})$. We call $\phi_j^{(i)}$ the SHAP value of input feature j for sample i. For more details of SHAP, please refer to [7].

Note that SHAP is a local method, which explains a model prediction based on each individual sample input. Thus, in this study, the same input feature has different SHAP values for different distribution feeders. By examining these SHAP values, we can discover which features have a significant contribution to the prediction output and how a input feature' contribution varies among different feeders.

In this paper, to interpret a prediction model, we calculate the SHAP value of all input features for every feeder in the dataset. Since each model is trained 10 times, the average SHAP value is reported as the final result.



(a) Histogram of # of customers. (b) Histogram of DF.

Figure 1: Overview of feeders and DFs in the dataset.

III. Descriptions of Real-World Distribution Feeders and Input Features for DF Prediction

In this section, we first describe the distribution feeders used in the case study and summarize the statistics for their DFs. Then, we describe the input features used to predict DFs.

The case study covers 3,952 distribution feeders managed by Southern California Edison. In total, these feeders serve over 4,000,000 customers. The histogram of number of customers for the feeders is shown in Fig. 1a. Using one year of hourly kWh readings of customers in 2015, we calculate the DFs of all distribution feeders according to equation (1). The histogram of DF is shown in Fig. 1b.

We collect various types of input features to predict DFs of distribution feeders. The input features can be categorized into three classes: feeder characteristics, customer demographic and socioeconomic conditions, and environmental factors. The input features are summarized in Table II. The sources of these input features are provided below.

1) Feeder Characteristics: Input features in this class represent the properties of the distribution feeder, which include number of customers, customer type, and the size and penetration rate of solar PV systems. These information is provided by Southern California Edison.

2) Demographic and Socioeconomic Conditions: Input features in this category are collected from the National Historical Geographic Information System (NHGIS) [10]. NHGIS organizes customers' data by census block groups (CBGs) instead of feeders. Thus, we derive the input feature values of each distribution feeder by matching the feeder's service area to the geographic locations of CBGs.

3) Environmental Factors: The California Energy Commission provides the climate zone information for each zip code [11]. By mapping the distribution feeders' locations to zip codes, we can obtain the climate zone information of each feeder. The weather data is collected from the National Centers for Environmental Information [12], which organizes weather data by weather stations. By mapping the feeder locations to weather stations, we can obtain the weather data for each feeder. The elevation of distribution feeders are collected from U.S. Geological Survey by queries using feeder locations.

IV. DF Prediction Performance and Interpretation of the Machine Learning Model

In this section, we first present the DF prediction performance of different machine learning models. Then, we quan-

TABLE II: Summary of Input Features

Class	Feature Type	Feature Description
Feeder Characteristics	No. of Customers	Number of customers in each feeder.
	Customer Type	Ratio of residential customers, ratio of commercial customers.
	Solar PV	Ratio of customers with solar PV and average solar PV size of commercial and residential customers, respectively.
Demographic and Socioeconomic Conditions	Age	Average age, ratio of population in 4 groups: child age (5 years), school age (6 17 years), work age (18 61 years), retired age (62 years).
	Education	Ratio of population in 4 educational levels: lower than college, less than 4 years' college, bachelor's degree, higher than bachelor's degree.
	Average Room No.	Average No. of rooms of a housing unit.
	Annual Income	Average household income, ratio of population in 3 income levels: $34,999, $35,000 $149,999, $150,000.
	Population	Population of each feeder's CBG.
	Occupancy Ratio	Occupancy ratio of housing units.
	Child Family Ratio	Ratio of families with children.
	Employment	Ratio of population in 4 conditions: employed, unemployed, army, not in labor.
Environmental Factors	Climate Zone	Building climate zone of California.
	Weather	Annual avg. of daily max. and min. temperature; annual highest, lowest, and avg. temperature; No. of days with max. temperature 90°F, 70°F, and 32°F; No. of days with min. temperature 32°F and 0°F; heating degree days; cooling degree days.
	Elevation	Elevation of feeders' service area.

tify the features' importance in determining feeder DF. Lastly, we analyze how different features affect the DF prediction and provide more insights into how DF is determined. The case study is conducted in Python on an Oracle-Sun workstation with 2.3 GHz Intel Xeon CPUs and 128 GB of RAM.

A. Prediction Performance of Machine Learning Models

The MSEs of 6 machine learning models on the test dataset are shown in Fig. 2. Each model is trained 10 times with the tuned hyperparameter setting and the MSEs are represented by the box plot. The red bar represents the median value, and the green diamond marker represents the mean value. The variance of DF in the test dataset is 0.22811. The MSE of the benchmark linear regression model is 0.13445. As shown in the figure, all 6 supervised machine learning models yield more accurate DF prediction results than the linear regression model. Among all tested models, random forest has the lowest average MSE. The figure also shows that pruning improves the accuracy of FNN and FNN+dropout models. FNN+dropout+pruning and GBT have a similar level of performance. All 6 machine learning models take less than 1 second to predict the DFs of the 3,952 feeders. Since DF prediction is often conducted as part of the distribution system planing process, the model training can be done off-line.

B. Importance of Different Feature Types

We calculate the SHAP values of all input features and samples for the random forest model, which yields the best

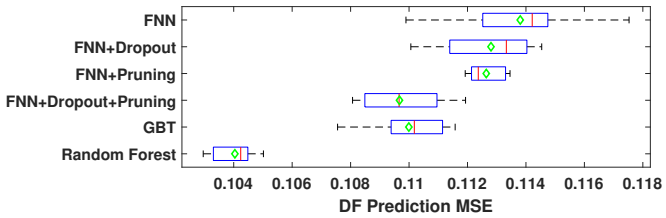


Figure 2: Box plot of the DF prediction MSEs of 6 models.

prediction results. We then derive the feature type importance as follows. First, for each feeder, we sum up the SHAP values of input features by feature type. Then, for each feature type, we calculate the average absolute value of the SHAP values over all feeders, which quantifies the importance of each feature type in determining DFs. Fig. 3 shows the importance of all feature types, ranked from the highest to the lowest. The most influential feature types are customer type, weather, solar PV, climate zone, and number of customers.

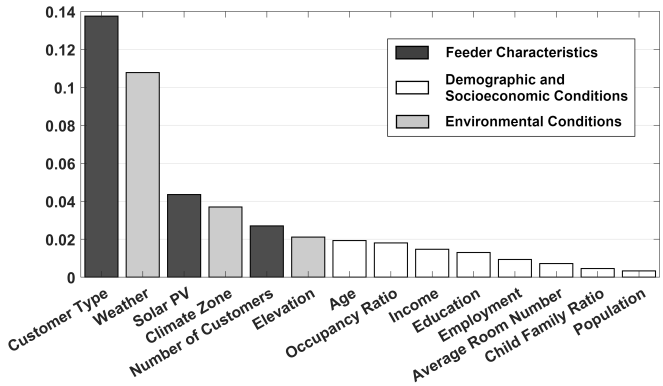


Figure 3: Feature importance of the random forest model

C. Impacts of Input Features on DF Prediction

For the random forest model, we select a few features with high importance and analyze their effects on DF prediction. To do so, we plot the SHAP values of a feature vs. the values of the feature for all distribution feeders in Fig 4. In the subfigures, each circle represents a feeder, and we can see how a feature's contribution to DF (i.e., the SHAP value) changes when the feature's value changes. To demonstrate the interactions between features, we color the circles by the ratio of residential customers in some subfigures.

1) The Impacts of Customer Type: In the testing feeders, all customers are either residential or commercial. As shown in Fig. 4a, feeders with higher ratio of residential customers tend to have higher DFs. This phenomena can be explained as follows. The electricity usage patterns of commercial buildings are less diversified because their demands usually follow normal business schedules. For example, restaurants, department stores, and cinemas often have similar operation hours. In comparison, residential customers often have drastically different electricity usage patterns due to the randomness of