# Machine Learning Applications for Predicting Renewable Energy Project Success in the PJM Interconnection Queue

Natalie Avida

## ABSTRACT

In recent years, the deployment of renewable energy has grown significantly, with both the investment and proposed capacity exceeding expectations. This rapid expansion has caused interconnection queues to balloon, doubling wait times and significantly reducing the probability that a project will actually reach operations. The literature generally agrees that the interconnection process requires further reform, and uncertainty is detrimental to renewable energy profitability and expansion. In this thesis, I use machine learning classification techniques to mitigate the uncertainty of a proposed energy project waiting in the interconnection queue. I compared logistic regression with standard regularization, logistic regression with elastic net regularization, and decision tree classification. Each of these models were trained on two datasets, one of which had all basic information about each project, renewable incentive policy information, and renewable resource quality (queue_basic), and one that had all of this information as well as data with the cost of interconnection for each project (queue_costs). The decision tree classifier performed best, with an accuracy of 80.0% for queue_basic and 90.6% for queue_costs, with the most important determinants being the year the project was proposed, the cost of interconnection, and MW output of the plant. Overall, this thesis demonstrates the applicability of machine learning to reducing renewable energy project uncertainty in the interconnection queue, and yields valuable insights about what aspects of a project determine its ability to execute an interconnection agreement.

## KEYWORDS

Interconnection queues, renewable energy, electricity, PJM, machine learning, logistic regression, decision tree classification

# INTRODUCTION

In recent years, the energy landscape in the United States has changed dramatically, with the vast majority percentage of proposed capacity being from renewable sources. Currently, the primary energy sources are still petroleum and natural gas, but renewables have made up an increasing proportion over the past few years. The source of energy production has major implications for climate change. In the United States, almost 30 percent of global warming emissions come from the electricity generation sector (EIA, 2017). Comparatively, renewable energy sources produce almost no global warming emissions, even when including the full life cycle of clean energy technology. For example, a 2009 analysis found that if 25 percent of the energy produced in 2025 was produced by renewable sources, $CO_2$ emissions would go down by over 275 million tons annually (UCS, 2009). Renewable energy development has scaled up significantly over the past few years, at a pace that has exceeded expectations. In 2022, investments in low-carbon energy "reached parity" with the amount of capital devoted towards expanding fossil fuels. Despite supply chain disruptions and economic downturn, the level of investment in the energy transition in 2022 grew by 31% (Catsaros, 2023). This trend is expected to continue if not grow in the coming years.

However, investment into a renewable energy development does not always mean that the project will be realized. Currently, there are over 2000 GW of capacity stuck in interconnection queues, which are lists of potential energy projects waiting to undergo a series of studies before finding out whether or not they will be able to connect to the electrical grid, and at what cost. The majority of these projects will never be built (Lenor, 2023). Requests have also increased steadily over the years. Renewables make up almost 95% of the projects in the queue, with solar and battery storage projects comprising over 80% of proposed capacity entering the queues in 2023 (Rand et al., 2024). Only 21% of the projects seeking interconnection agreements in the span of 2000 to 2017 have actually been built as of the end of 2022, and wait times have increased from less than two years to over four years (Rand et al., 2024). The process of receiving interconnection is intensive—after entering the queue, projects must undergo a series of interconnection studies which identify any new required transmission equipment or upgrades needed, which then assign the cost of those upgrades. If all the requirements are met, a contract between the ISO or utility and the project developer is created, which is called an interconnection

agreement. However, the majority of projects, many of which already have millions of dollars worth of development, are withdrawn from the queue prior to this point (Rand et al., 2024).

The PJM interconnection queue is the second largest in the United States, with over 300 GW. PJM is a regional transmission operator serving over 65 million people. It has the largest stand-alone solar capacity, with a proposed 170 GW—34.1 percent of the national total. In the past few months, PJM has faced significant challenges in accommodating this massive surge in renewable energy projects, particularly in terms of transmission infrastructure (Amman, 2023). Efficiently transmitting and distributing the generated electricity is critical to realize the full potential of these clean energy sources. A robust and well-maintained transmission system is essential to ensuring the reliability and stability of the grid while preventing congestion and bottlenecks. Additionally, having projects remaining stagnant in the queue for years is a major waste of resources, both for PJM and developers. Predicting a project's likelihood of being approved for interconnection, as well as the timeframe within which it would happen, can prevent these resources from being wasted.

## BACKGROUND

### History of Electricity Regulation and Regional Transmission Organizations

The generation and transmission of electricity in the United States has changed dramatically since the creation of the electric grid in the late 19th century. Electricity generation has been determined by a combination of policy and market forces, beginning with rapid vertical integration in the late 1800s, followed by a period of the regulated monopoly, and most recently deregulation (Nudell et al., 2018). Prior to the 1970s, electricity generation was primarily controlled by government-approved utilities. However, in response to the 1973 oil crisis, the US Congress passed the National Energy Act of 1978, which included a key statute: the Public Utility Regulatory Policies Act (PURPA). PURPA opened the market to non-utility generators or independent power producers who were able to produce power at a lower cost than utilities (Handmaker, 1989). PURPA was followed by the Energy Policy Act of 1992, which mandated that a utility provide transmission access to large buyers and merchant generators (Nudell et al., 2018). To manage the new need for transmission, in 1999 the Federal Energy Regulatory

Commission (FERC) created Regional Transmission Organizations (RTOs). RTOs are electric power transmission system operators which regulate, manage, and monitor multi-state electric grids. The establishment of RTOs has had a major impact on power generation and transmission in the United States. Since RTOs rarely own transmission facilities or power generation facilities, FERC believes that they have the unique opportunity to prevent bias in selecting generation sources, use market-based approaches to solve congestion issues, and improve reliability through the planning and operation of regional transmission infrastructure (Porter, 2002). By overseeing grid coordination and flexibility, RTOs currently do and will continue to play a pivotal role in advancing the energy transition.

**Overview of the Interconnection Queue Process**

One of the critical roles of Regional Transmission Organizations (RTOs) involves managing interconnection queues through the conduct of interconnection reviews, a process that is pivotal in integrating new energy projects into the grid. While every region has its own set of processes, every potential generator is required to submit a detailed application, pay a deposit, and show that it will be able at some point to have control of the site through land-use permits from the outset. The costs from the beginning of the process are not trivial—the non-refundable deposit is $10,000 for PJM, and the study deposits can cost hundreds of thousands depending on the size of the project (Egan, 2015). Once these initial steps have been completed, the project developer works with the grid operator to complete a series of studies that assess the project's impact on the electrical grid. The typical studies conducted in different phases are a feasibility study, a system impact study, and a facilities study. The feasibility study examines whether the energy project would require transmission updates to connect with the grid. The system impact study requires a higher level of detail from the potential generator, as it examines grid impacts in more detail. At this point, the developer can still change some details about the project. The facilities study is the most rigorous, estimating in the highest level of detail the costs of all aspects of the facilities needed to connect the project to the grid, such as the necessary equipment, engineering, and construction. At this stage, the project's design and details must be relatively finalized (American Clean Power, 2023). Overall, a project's path through the interconnection process can take up to four years (PJM's average wait time is 24.4 months), and

the vast majority of projects in the interconnection queue never reach commercial operations (Rand et al., 2024). Developers are able to withdraw at any point during the process, with many withdrawing later in the process after years of work and investment.

**PJM Interconnection & the PJM Interconnection Queue**

As the coordinator of one of the largest electricity markets and one of the oldest RTOs in North America, PJM has set a precedent in several aspects of interconnection processes, particularly with its innovative approaches to managing grid reliability, its comprehensive market structure, and its transparent and efficient interconnection procedures. Established in 1927 as the Pennsylvania-New Jersey-Maryland Interconnection, it originally served as a power pool designed to coordinate the wholesale electricity market in the mid-Atlantic region (PJM, 2023). Over the years, PJM expanded its role and became one of the nation's first RTOs in 1997. As an RTO, PJM plays a crucial role in managing the transmission of electricity across a multi-state region, ensuring grid reliability, and overseeing competitive wholesale electricity markets. It has grown to encompass 13 states and the District of Columbia, serving as a model for grid management and market operation in the United States. Acting as a third party, PJM manages the largest competitive wholesale electricity market in the world and controls the operation of an electricity grid which reaches over 65 million people (PJM, 2023). In order to manage its high volume of projects constantly being added and studied in the interconnection queue, PJM has instituted several reforms to expedite the interconnection queue wait times. One significant reform that PJM has implemented is shifting from a first-come/first-served serial interconnection study process to a first-ready/first-served cluster study model (PJM, 2023). This approach addresses critical issues such as delays and cost allocation challenges in the interconnection process. By grouping projects in clusters, costs for necessary network upgrades are shared among all projects within the cluster, rather than the first project that needs upgrading bearing a disproportionate cost (Cannon & Wiseman, 2022). This new model aims to support only viable projects, reducing the number of withdrawal and improving the energy generation integration process.

Despite these advancements, the transition toward renewable energy within PJM is not progressing rapidly enough to mitigate the most severe effects of climate change. PJM’s current

market structure does not adequately integrate or account for the requirements and advantages of clean energy. In fact, PJM scored last on with a D-minus on Advanced Energy United's (AEU) 2024 Generator Interconnection Scorecard. AEU cited PJM's poor use of regional transmission planning and their lack of utilization of interconnection alternatives (Wilson et al., 2024). PJM is currently considering changes to its capacity markets, with the potential implementation of a Forward Clean Energy Market (FCEM) and an Integrated Clean Capacity Market (ICCM). Dedicated renewable energy markets could attract more investment into renewable energy projects by providing more certainty around the financial returns of these projects, potentially reducing wait times in interconnection queues (Glazer et al., 2022). However, until these or other policy shifts come into fruition, understanding the current interconnection queue process and its limitations is essential. Identifying the patterns and predictors of success under the status quo is key for providing a data-driven foundation for future policy considerations and improvements.

## RESEARCH FRAMEWORK

### Current Research on Transmission in the United States

Currently, the vast majority of the projects in need of interconnection into the grid are renewable energy projects. Importantly, across the five major Independent System Operators (ISOs) in the United States, only 19% of projects proposed from 2000-2018 have reached commercial operations as of the end of 2023 (Rand et al., 2024). Studies of the current state of interconnection in the United States overwhelmingly conclude that more transmission and comprehensive reform of the interconnection process is necessary. The majority of the literature reaches these conclusions primarily from a historical or policy-based perspective. Some scholars suggest that part of the transmission bottleneck is caused by "entrenched interests" of energy companies that have an incentive to maintain the status quo (Cantafio & Nowak, 2021). Other articles have analyzed the role of market dynamics and pricing in motivating building new transmission projects. Numerous researchers have determined, based on policy analysis and stakeholder investigations, that a significant impediment to the expansion of transmission infrastructure lies in contentious discussions surrounding the financing and governance of new transmission initiatives between different regions and governments (Cifor et al., 2014). Other

authors also discuss how the lack of interconnection presents a barrier to the energy transition and renewables development (Mays, 2023). However, while both the problem and the solution is thoroughly discussed in literature, there is a gap surrounding project by project impacts of transmission wait times, and solutions for renewable developers in the short-term.

**Impact of Uncertainty on Renewable Energy Projects**

A major issue that interconnection queues create in a renewable energy project is uncertainty. Every potential barrier to a project reaching commercial operations adds an additional layer of uncertainty, which means a higher probability that resources are inefficiently allocated or the project may never even exist. Assessing the precise effects of ambiguity on a project-specific level is challenging. However, researchers have explored the repercussions of various uncertain aspects in renewable energy development, such as storage limitations and intermittency, while also proposing potential solutions. Some of these solutions involve spatial modeling, as well as other technical approaches (Srinivasan et al., 2023). Another major source of uncertainty in renewable energy development is policy uncertainty. A key finding from this area of research is that unclear policy futures have a significant negative impact on renewable energy development (Khan & Su, 2022). Overall, a clear conclusion can be drawn—uncertainty is harmful for renewable energy development. In the context of interconnection queues, Yang, et al. (2024) found that mitigating uncertainty in the interconnection queue through proposed policy reforms such as evaluating clusters of generators at once would increase renewable energy deployment significantly. While the literature has presented policy options in a goal to reduce uncertainty throughout renewables development and in the interconnection queue, there has not been a breadth of literature attempting to mitigate it under the status quo. In this thesis, my objective is to explore the impact of resource loss due to the insecurity associated with successfully executing an interconnection agreement.

**Implementation of Data Science Classification in Renewable Energy Literature**

Machine learning classification techniques have been utilized to improve decision-making and predictive accuracy within the realms of renewable energy management and power

system forecasting with increased frequency and applicability in recent years. Data science approaches such as logistic regression have become increasingly prevalent in climate governance and energy planning literature as data-driven processes to support decision making and resource allocation. In their exploration of renewable energy adoption in Semarang, Indonesia, Ulkhaq et al. (2018) leveraged logistic regression analysis to identify key factors influencing consumer intentions towards embracing renewable energy solutions. Their study aims to show how specific determinants forecast the likelihood of consumer transition to renewable energy sources, offering valuable insights for policymakers aiming to enhance renewable energy uptake in the region. Similarly, Liu et al. (2022) utilized multivariate logistic regression to estimate the probability of extremely high and low electricity prices in the day-ahead Australian National Electricity Market. The model performed with high accuracy and was also useful in identifying the relative importance of the different variables, strengthening electricity price forecasting theories and overall understanding of extreme price dynamics. These studies, among others, demonstrate the current value of logistic regression models for both developing an efficient way to predict outcomes in the energy space and identifying the key determinants of these outcomes. These results help stakeholders derive actionable insights from large amounts of data, as well as develop more effective interventions tailored to the specific drivers of the specific outcomes they aim to achieve.

While the application of logistic regression to energy management classification problems has been demonstrated in the literature, use of decision trees for similar problems has received considerably less attention. Primarily, decision trees have been used for optimization and site selection purposes. Shorabeh, et al. (2022) compared a decision tree to a particle swarm optimization (PSO) algorithm to determine which method would better detect potential areas for solar energy sites in Iran. The decision tree algorithm outperformed significantly, achieving a prediction rate of 0.29 for identifying high potential solar development sites, compared with the Particle Swarm Optimization (PSO) algorithm, which had a much lower accuracy of 0.13 for the same category. Within the healthcare sphere, decision trees have been used for classification problems. Khempila and Boonjing (2010) compared the performance of logistic regression, decision trees, and artificial neural networks to predict heart disease incidence in patients. While neural networks achieved the highest classification accuracy, their black box nature and validation difficulties make the results more difficult to interpret. These limitations affect the

practical utility of neural networks in energy systems planning, particularly in discerning the significance of different determinants and understanding the implications for policy development. However, the comparable performance of the decision tree model, as well as the interpretability and simplicity of decision trees, demonstrates a gap in the literature in classification for clean energy management.

This thesis builds on the contextual framework provided by these three distinct areas of research, as well as others. Using machine learning classification to mitigate uncertainty in renewable energy development, specifically with regards to interconnection queues, presents a unique opportunity to support developers and policy development as well prevent suboptimal resource allocation. The following section will explain the methods employed, harnessing data science techniques to explore and address the challenges posed by the interconnection queue and potentially offer clarity to enhance project success and sustainability in the evolving energy landscape.

## METHODS

### Composite Dataset Formation

*Data collection and feature creation*

To assemble the datasets that the final model was developed on, I gathered and put together a wide variety of data including basic information about the projects, policy features, and renewable resource quality. The primary key of the datasets are power plants that have at some point been in PJM's interconnection queue. The basic information about the proposed projects was obtained through the Lawrence Berkeley National Laboratory (LBNL) website and through PJM's website. The data on PJM's website is obtained directly from the requests for interconnection processed by the TSO (PJM, 2024). The data downloaded from LBNL was compiled as part of "Queued Up", a study on the characteristics of energy projects that are currently seeking or have sought interconnection agreement as of the end of 2023. The data for the study was collected from interconnection queues for seven ISOs / RTOs and thirty-five utilities for projects through 2022. LBNL then standardized and cleaned the data before

publishing it (Rand, 2024). The full sample includes 29,154 projects, but I only utilized the data in the PJM region.

The PJM dataset has many more columns than LBNL's, as it also includes links to the studies that were conducted through the interconnection queue process. These columns were not relevant to my classification. The key advantage to LBNL's dataset is that it has already cleaned the interconnection agreement status, which was my target variable. I merged the two datasets to ensure data integrity and consistency in my predictions. This merge creates the base of one of the two datasets that the model is trained on, referred to going forward as queue_basic.

I also retrieved LBNL's PJM cost of interconnection dataset, as cost is a key consideration for utilities and developers when choosing whether or not to begin construction on a project. This dataset is much shorter than the other two datasets, so I performed two classifications, one on a dataset with the cost values merged in and one without so that I could identify which classification performs better—one trained on a larger dataset or one trained on a dataset with cost features. The LBNL cost dataset was compiled for a policy brief titled "Interconnection Cost Analysis in the PJM Territory". The brief analyzed interconnection cost data from 1,127 projects that were assessed in interconnection studies between 2000 and 2022. For each project that was assessed, LBNL calculated the point of interconnection cost per kW, the network cost per kW, and the total cost per kW. Data for 1,027 projects was collected from PJM's website, and data for 55 additional projects was collected in 2018 and has since been removed from PJM's online system (Seel et al., 2022). This dataset is significantly more limited than the projects dataset without costs, as the brief only focuses on new generation plants, applications that have completed a feasibility study, and not projects that have withdrawn and applied again later. Due to these limitations, all projects that entered the queue after March 2021 were not included in the dataset (Seel et al., 2022). Given that this dataset is significantly shorter than queue_basic, all model selection and testing was performed separately and concurrently on a version of a dataset built off of queue_basic, and a version of queue_basic with the cost dataset merged in, referred to as queue_costs. Only projects in queue_basic that have cost data were included in that dataset.

Another key aspect of the datasets was policy data on the renewable incentives for each state and utility in PJM, as I hypothesized that a state's renewable energy policies would enhance the accuracy of the model. I manually created two policy datasets in Excel to describe the policy

incentives relevant to each project based on the state and utility where the proposed project was located. I obtained this information from the Smart Electric Power Alliance's (SEPA) Utility Carbon-Reduction Tracker, which features interactive maps that display which US states have instituted a mandatory 100% renewable or clean energy standard and which utilities have instituted various carbon reduction goals (SEPA, 2024). SEPA compiles the information provided in the Utility Carbon-Reduction Tracker Sourced directly from documents issued by individual utilities, utility parents, generation and transmission cooperatives (G&Ts), and state governments (SEPA, 2024). The state level policy dataset I created includes every state in PJM, and includes the state and whether or not they have implemented a 100% renewable energy or clean energy standard (represented by a 1 or a 0). The utility level policy dataset I created includes every utility represented in PJM and which type of renewable energy target it has adopted (represented by a 1 or a 0): 100% carbon-free / renewable energy, net negative, net-zero or carbon neutral, or partial reduction. If the utility does not have a renewable energy goal it has a 0 for all categories. After creating these datasets, I merged them both into queue_basic and queue_costs.

I also collected renewable resource quality data, specifically solar and wind resource quality data, to provide crucial insights into the potential efficiency and productivity of renewable energy projects within the interconnection queue. Both solar and wind data was collected from the National Renewable Energy Laboratory (NREL). The solar PV supply curve dataset was downloaded from the website and includes capacity factor, irradiance, MW capacity, and distance to transmission lines for 55,534 latitude-longitude (lat-long) coordinates across the United States (Lopez et al., 2021). These data are taken from NREL's National Solar Radiation Data Base (NSRDB), which was developed using satellite data and a 2-step physics based model known as the PSM (Sengupta et al., 2017). The wind supply curve data was downloaded from one of NREL's Wind Integration National Dataset (WIND) Toolkits and includes wind speed, capacity factor, and fraction of usable area for 126,691 lat-long coordinates across the United States and offshore (Draxl et al., 2022). This dataset was created by NREL primarily using the Weather Research and Forecasting modeling framework, which is maintained by the U.S. National Center for Atmospheric Research (Hodge, 2015).

In order to incorporate the resource availability data into the primary datasets, I had to assign a latitude and longitude value to every project to find the nearest resource quality

datapoint. Lat long coordinates were provided directly to me from the LBNL, from a dataset that was scraped from the PJM queue map. For projects not featured in this dataset, I utilized county spatial data to match the proposed power plant's county with a lat-long coordinate at the county's center in order to approximate the project's location. Once every project was assigned a latitude and longitude value, I employed geospatial analysis techniques, specifically cKDTree, to pinpoint the nearest coordinate pair within both solar and wind resource datasets. cKDTree is a data structure from the SciPy library used for efficient spatial indexing and nearest-neighbor queries (SciPy, 2024). I defined two functions for the solar and wind respectively that converted the geographical coordinates into NumPy arrays for both the target locations and renewable energy sources then created cKDTree spatial indices for these points. These functions allowed me to automatically identify the nearest solar and wind points to each target location. The functions then appended the relevant resource quality data, as well as the distance between the target coordinates and the renewable coordinates, to my target dataframes. The result was seven additional features for both datasets that provide insight on the potential of nearby renewable energy resources. Ultimately, after compiling all of these datasets and selecting relevant features and prior to one-hot encoding, queue_basic had 22 columns which serve as features and queue_costs had 25 columns which serve as features.

*Data cleaning and feature engineering*

After creating all the features, I cleaned the datasets to ensure their accuracy, efficiency, and reliability for data analysis and modeling. Cleaning data refers to removing and correcting inconsistencies, errors, and missing values. One of the most important steps in cleaning queue_basic and queue_costs was creating the target variable, which was the Interconnection Agreement (IA) status. In reality, there are many different stages of the interconnection queue, but my aim was only to determine if a project would eventually execute an IA or withdraw from the queue. Consequently, I consolidated the twelve different statuses in the dataframe column labeled 'IA_Status_Clean' into three categories: "In Progress", "Executed", and "Withdrawn" (Table 1).

**Table 1: Mapping of interconnection categories to simplified labels.** The left column shows the simplified statuses that were mapped to the original labels which represent a more granular view of the status of the project. However, for the purposes of my model, only three labels were necessary.

| Simplified Label | Original Label |
|---|---|
| In Progress | Active |
| | Confirmed |
| | Suspended |
| | Engineering and Procurement |
| Withdrawn | Annulled |
| | Cancelled |
| | Deactivated |
| | Retracted |
| | Withdrawn |
| Executed | In Service |
| | Under Construction |
| | Partially in Service - Under Construction |

A critical part of the data cleaning process was removing NaN (not a number) values from the dataset. NaN values occur when a value does not exist in a dataset. Machine learning algorithms cannot be run on datasets that have NaN values. There were not many NaN values in my dataset, so there were only a few steps I had to take to clean it. For the NaN values in the latitude and longitude columns, I dropped the rows where these occurred, as I determined that these projects were too early in their development that there was not enough relevant information to impute these values. There were 85 rows with NaN values in the latitude and longitude columns in queue_basic and 0 in queue_costs. I filled the NaN values in the Maximum Facility Output (MFO) column with the values from the MW capacity column as there was significant overlap between these columns already.

After ensuring that all NaN values were removed and the data was clean, I began the process of feature engineering, which transforms raw features into more informative features for modeling. Feature engineering helps to capture more knowledge that is not explicitly outlined in the dataset and model non-linear relationships with linear models (Crouch et al., 2023). A core part of my feature engineering process was using one-hot encoding to transform the categorical variables into numerical variables, as linear machine learning models cannot interpret categorical variables. I used the OneHotEncoder function from scikit-learn, a free software machine learning library, to map all of the categorical variables, specifically State, County, Transmission Owner, Energy Type, and Interconnection Agreement Status into binary vectors, allowing me to use these variables as features for the models.

In addition to one-hot encoding, I executed feature engineering to create additional features, including the creation of polynomial features in the second and third degrees for key numerical variables such as MW output, the nearest solar and wind capacity factors, the nearest solar irradiance and the nearest wind speed for queue_basic, along with costs of interconnection for queue_costs. I also introduced a "num_projects_proposed" feature to quantify the volume of project proposals for every year, offering insights into the competitive environment at the time each project was proposed. To approximate infrastructure proximity, I combined the distance between the project and the solar resource coordinate with the distance between the solar resource coordinate and the transmission line. This process may have inaccuracies as these distances do not take into account directional information, but I knew regularization in the modeling process would offset any impact from potentially non-informative features. Additionally, I developed three interaction terms to enhance our analysis. The first term, named "capacity_accessibility," was generated by multiplying the MW output with the distance to the nearest transmission line. For the other two terms, I created separate columns that specifically account for the capacity factor of solar and wind projects, applying these metrics only to projects that incorporate solar or wind components, respectively. Ultimately, I created 14 new potential features for queue_basic and 20 new potential features for queue_costs, helping improve the predictive accuracy of the model. The final versions of queue_basic and queue_costs are linked in the appendix.

*Exploratory data analysis*

Exploratory data analysis (EDA) is a critical part of developing a model, where visualizations are developed to understand trends in the data. I performed EDA on both queue_basic and queue_costs in order to understand some of the underlying trends within the data. Importantly, I confirmed that the majority of the projects in the datasets were renewable energy projects (Figure 1).

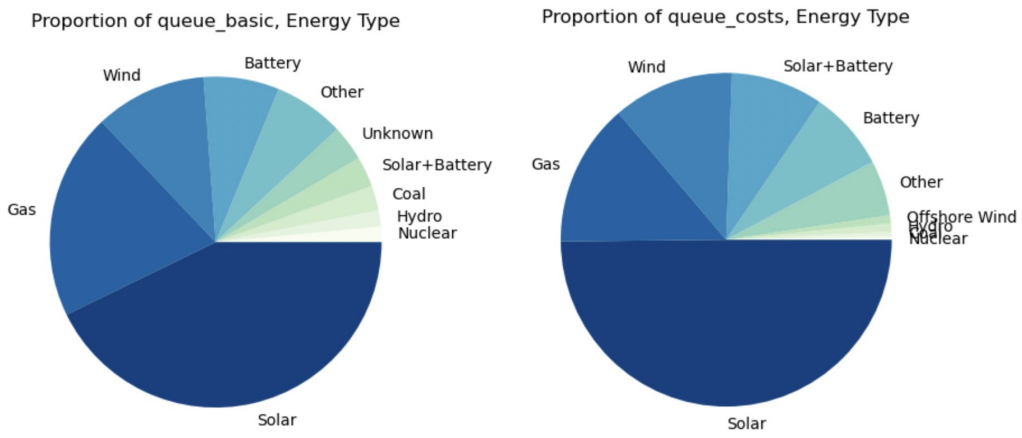


**Figure 1: Proportion of data based on energy type, queue_basic (left) and queue_costs (right).** Two pie charts with the sections indicating the proportion of different energy types in the two datasets. The sections are labeled and the different colors indicate different types of energy.

These charts confirm that the majority of the data are renewable energy projects, with solar being the largest category and wind also representing a large share.

Additionally, I analyzed the change in the number of entries to the interconnection queue over time. The number of proposals submitted to the PJM interconnection queue in a given year changes significantly over time (Figure 2).

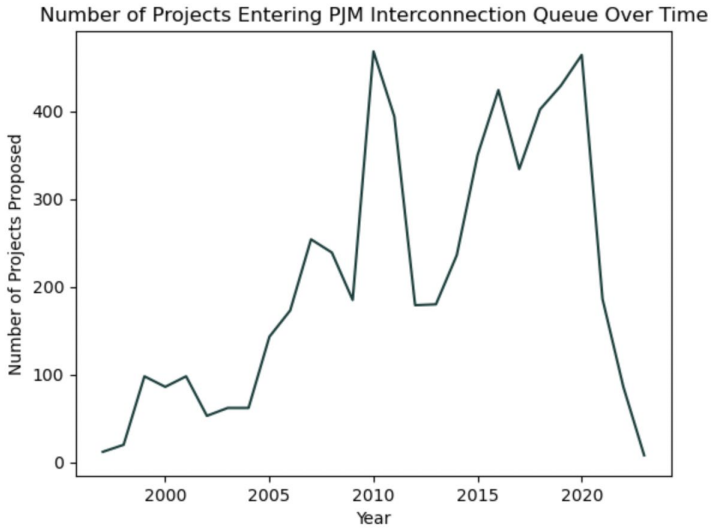


**Figure 2: Number of projects proposed per year, queue_basic.** Line chart showing the number of projects entering the queue at a given year. The y-axis is the number of projects proposed, and the x-axis is the year.

Despite a large dip in the early 2010s, the overall quantity of projects entering the queue has increased over time. The dip in the number of projects entering the queue at the end is due to the dataset only taking into account some of the proposed projects in the years since 2021 as the source data has not been updated.

In addition to a wide variation in the number of projects entering the queue over time, there is also large variation in the outcome with these projects over time, particularly in queue_costs. The balance between withdrawn projects and executed projects is relatively equal in queue_basic, but changes drastically over time in queue_costs (Figure 3). This analysis informed the decision, described later, to also test model performance on a version of the datasets without temporal features.

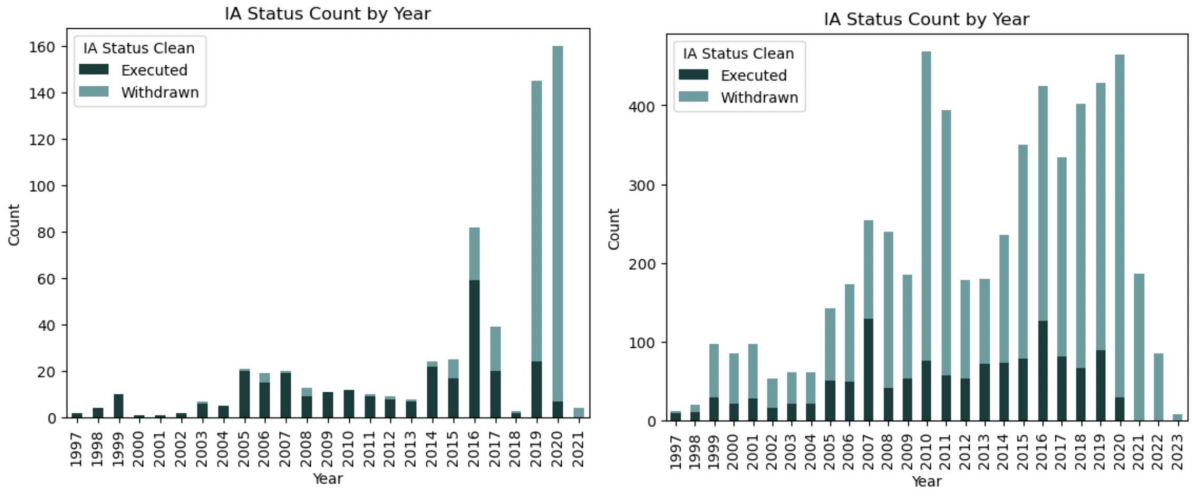


**Figure 3: Outcome of projects proposed each year, queue_basic (left) and queue_costs (right).** Two stacked bar charts showing the ratio between the number of projects that ultimately executed interconnection agreements and the number of projects that ultimately withdrew from the queue. The y-axis represents the count, and the x-axis represents the year.

The total cost of interconnection also varied significantly over energy type (Figure 4).

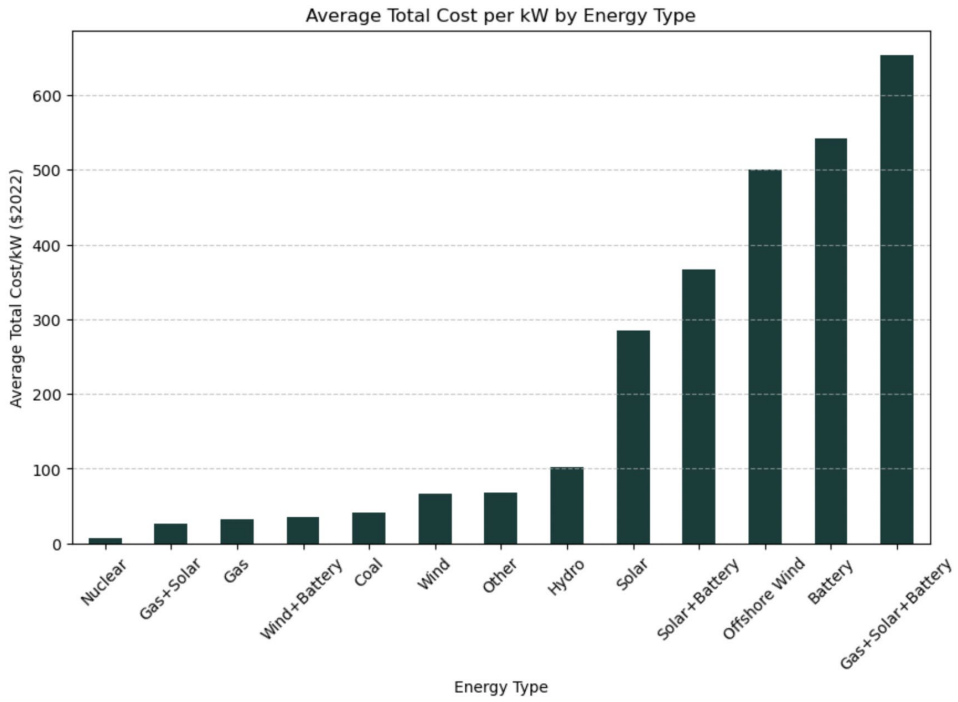


**Figure 4: Average total cost/kW by energy type.** Bar chart showing the average total cost of interconnection in $/kW as of 2022 for each energy type, with the cost increasing towards the right. The y-axis is $/kW and the x-axis is the type of energy.

This analysis is in line with general trends in energy costs, with more complex multi-type projects being the most expensive, along with battery storage projects and offshore wind, indicating data integrity. Ultimately, the exploratory data analysis informed subsequent modeling and analysis decisions, specifically regarding importance of various features and creating subsets of the data.

## Classification Model Implementation

*Logistic regression*

Logistic regression (LR) is a unique linear regression model that is usually used to predict the value of a binary dependent variable. LR starts from a linear equation, but uses log-odds in order to constrain the output between 0 and 1 (Khemphila & Boonjing, 2010). An LR model for n independent variables, known as the LR loss function, is written as

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots++\beta_n x_n)}}$$

LR is described as a linear model as the result of the natural log of the ratio of $P(Y = 1)$ to $1 - P(Y = 1)$ or $P(Y = 0)$ is a linear model, written as

$$f(x) = ln\left(\frac{P(Y=1)}{P(Y=0)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Logistic regression with regularization restricts complexity by adding a penalty on the size of the coefficients to the loss function. Regularization is a tool to tune bias and variance and improve model performance. There are two types of regularization—L1 (Lasso) and L2 (Ridge). L1 regularization adds a penalty equal to the absolute value of the magnitude of coefficients, causing some coefficients to be shrunk to zero. As a result, L1 can be used to perform feature selection:

$$\text{Loss Function} + \lambda \sum_{i=1}^{n} |\beta_i|$$

L2 regularization adds a penalty equal to the square of the magnitude of coefficients, meaning no coefficients are shrunk to zero—all are reduced by the same factor. As a result, Ridge's feature selection is not interpretable:

$$\text{Loss Function} + \lambda \sum_{i=1}^{n} \beta_i^2$$

Elastic Net regularization combines both L1 and L2 regularization. Elastic Net is useful for dealing with highly correlated variables as it is very useful for grouping variables together and assigning them the same level of importance:

$$\text{Loss Function} = \lambda \left( \frac{1-\alpha}{2} \sum_{i=1}^{n} \beta_i^2 + \alpha \sum_{i=1}^{2} |\beta_i| \right)$$

In all of these formula, $\lambda$ is the regularization hyperparameter that controls the strength of the penalty. The value chosen for $\lambda$ is critical for preventing overfitting while maintaining the accuracy of the model. $\alpha$ is an additional hyperparameter used for Elastic Net regularization to balance the contribution of L1 and L2 regularization, with an $\alpha$ of 1 representing 100% Lasso regularization and an $\alpha$ of 0 representing 100% Ridge regularization (Friedman et al., 2010) (Crouch et al., 2023).

*Decision Tree Classifier*

Decision trees are non-parametric models, meaning that they do not assign coefficients to features but continuously split the data in a hierarchical structure until reaching the final outcomes. Specifically, decision tree classifiers (DTC) construct a decision tree to model the decision process and arrive at conclusions about the target variable based on input features. Decision trees start at the root node and split the data into internal nodes until reaching leaf nodes, which for DTCs, are either 1 or 0. At each node, the algorithm selects the best feature to split on, aiming to partition the data in a way that increases homogeneity within each subsequent node. The splits are determined by a specific criteria, which for classification purposes, is generally either Gini Impurity or information gain and entropy. Gini Impurity is a measure of how frequently a randomly chosen item from the dataset would be incorrectly labeled if it was randomly labeled either 1 or 0, represented mathematically by:

$$\text{Gini} = 1 - \sum_{i=1}^{n} p_i^2$$

In this formula, $p_i$ represents the fraction of data points with class $i$ in the dataset (Karabiber, 2024).

Information gain measures the change in entropy after a dataset is split by a decision tree. Entropy quantifies the randomness or variance of a dataset. The goal of a split is to reduce the overall entropy of the child nodes relative to the parent node. The entropy for a classification problem is measured by:

$$\text{Entropy} = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

Like the Gini Impurity formula, $p_i$ represents the probability of randomly choosing a data point of class $i$ from the dataset (Zhou, 2022). Both Gini Impurity and entropy affect how decision trees decide to split data at each node, potentially leading to different tree structures. In addition to these criterion, there are other key hyperparameters that are not directly included in a mathematical equation, but affect the performance of the classifier, namely the maximum depth of the tree, the minimum number of samples required to split an internal node, and the minimum number of samples required to be at a leaf node. The maximum depth of the tree controls how many times the tree splits the data, which when selected appropriately, can help balance bias and

variance. The minimum sample size required to split a node also is important for balancing accuracy and overfitting—a large sample size reduces variance, but increases overfitting. The minimum number of samples required to be at a leaf node also helps control overfitting by ensuring that leaves contain more than just a few samples. Selecting these hyperparameters is crucial for achieving model precision while maintaining its ability to generalize across unseen data.

*Implementation and Model Selection*

I implemented the LR model using the LogisticRegression class from the sklearn.linear_model module, a widely recognized library in the Python programming language for machine learning. The scikit-learn implementation provides a robust, flexible framework for logistic regression, including support for various regularization techniques. The optimal hyperparameters were selected through a grid search cross validation process using the GridSearchCV class from the sklearn.model_selection module. Grid search cross validation tests every hyperparameter combination on subsets of the input data and evaluates them on the remaining subset of the data in order to find which hyperparameters yields the highest accuracy.

I conducted a grid search for both standard regularization hyperparameters and elastic net hyperparameters on both queue_basic and queue_costs. For the standard regularization model, the parameter grid I evaluated included L1 and L2 regularization types, along with varying values of C (scikit-learn's regularization strength hyperparameter) across a logarithmic scale from 0.001 to 100. It is noteworthy that C represents the inverse of $\lambda$, the regularization strength, where smaller values indicate stronger regularization. For the elastic net model, I evaluated varying L1 to L2 ratios in increments of 0.25 from 0 to 1, along with the same values of C used for standard regularization excluding 0.001 given how computationally intensive this grid search was. The grid search determined that standard regularization had an overall higher accuracy than elastic net regularization for both queue_basic and queue_costs, with the optimal hyperparameters being an L1 penalty and a C value of 0.1 for queue_basic and an L1 penalty and a C value of 0.1 for queue_costs. A preference for L1 regularization for both of the datasets indicates a preference for a model that has feature selection, given that L1 regularization drives some coefficients down to zero. This outcome makes sense as there are many features in the dataset that are potentially irrelevant for the outcome, particularly as a result of the large number

of categorical variables. A C value of 0.1 indicates somewhat strong regularization, imposing a higher penalty on the magnitude of coefficient. I incorporated these hyperparameters into the final models, which I then trained and evaluated on their respective datasets.

I implemented the DTC model using the DecisionTreeClassifier class from the sklearn.tree module. The scikit-learn application is highly flexible and allows for the implementation of key hyperparameters. The optimal hyperparameters were selected through the same process that I used for the LR models, via the GridSearchCV class from the sklearn.model_selection module. For both queue_basic and queue_costs, I tested a variety of options for four key hyperparameters: the selection criterion, the tree's maximum depth, the minimum samples required to split a node, and the minimum number of samples required to be at a leaf node. For the criterion, I tested both Gini and entropy. For the maximum depth, I tested increments of ten ranging from 0 to 50. For the minimum samples required to split a node, I tested 2, 5, and 10, and for the minimum number of samples at a leaf node, I tested 1, 2, and 4.

The grid search revealed markedly similar optimal hyperparameters for queue_basic and queue_costs. For queue_basic, the grid search showed that the optimal configuration of hyperparameters was 'gini' as the criterion for feature selection, a maximum tree depth of 10, a minimum requirement of 1 sample for splitting a node and a minimum leaf sample size of 2. For queue_costs, the grid search concluded that the optimal hyperparameters was also 'gini' as the criterion for feature selection, a maximum tree depth of 10, a minimum requirement of 1 sample for splitting a node and a minimum leaf sample size of 5. These selections indicate a preference for a moderately complex model that avoids overfitting while still capturing essential patterns. I integrated these hyperparameters into the final models and subsequently trained and assessed them on their corresponding datasets.

The optimals models were trained and tested on three variations of both queue_basic and queue_costs—the full dataset, a subset with only renewable energy projects, and the full dataset without temporal variables.

## RESULTS

### Initial Model Selection and Performance

As described in the methods section, three different models underwent hyperparameter selection on both queue_basic and queue_costs—standard regularization logistic regression, elastic net logistic regression, and decision trees. I selected the hyperparameters based on accuracy, with the best cross-validation accuracy for each model shown in Table 2.

**Table 2: Best parameters and cross-validation accuracy for all models, queue_basic and queue_costs.** Table showing the best parameters and the best cross validation accuracy for each method and both datasets. Each row is a different method with the corresponding best parameters and best cross validation accuracy for both datasets.

| | queue_basic | | queue_costs | |
|---|---|---|---|---|
| **Method** | **Best parameters** | **Best cross validation accuracy** | **Best parameters** | **Best cross validation accuracy** |
| Standard Logistic Regression | penalty: l1, C: 0.1 | 79.02% | penalty: l1, C: 0.1 | 83.73% |
| Elastic Net Logistic Regression | l1 ratio: 1, C: 10 | 77.62% | l1 ratio: 0.5, C: 1 | 81.70% |
| Decision Tree Classifier | criterion: gini, max depth: 10, min samples leaf: 1, min samples spit: 2 | 79.16% | criterion: gini, max depth: 10, min samples leaf: 1, min samples spit: 5 | 84.85% |

I then fitted the selected hyperparameters to a training subset of the entire dataset (80% for queue_basic, 85% for queue_costs) and tested on the remaining subset. A comparison of the accuracy (ACC), precision (PRE), recall (REC), and negative predictive rate (NPR) are shown in Table 3 and Table 4 for queue_basic and queue_costs, respectively.

**Table 3: Performance metrics for queue_basic.** Table with percentage accuracy, precision, recall, and negative predictive rate for each type of model tested on queue_basic.

| Method | ACC | PRE | REC | NPR |
|---|---|---|---|---|
| Standard Logistic Regression | 79.1% | 94.4% | 81.3% | 28.6% |
| Elastic Net Logistic Regression | 78.6% | 92.4% | 82.0% | 33.2% |
| Decision Tree Classifier | 80.0% | 93.0% | 83.0% | 37.0% |

**Table 4: Performance metrics for queue_costs.** Table with percentage accuracy, precision, recall, and negative predictive rate for each type of model tested on queue_costs.

| Method | ACC | PRE | REC | NPR |
|---|---|---|---|---|
| Standard Logistic Regression | 83.3% | 81.1% | 87.8% | 86.0% |
| Elastic Net Logistic Regression | 81.3% | 81.1% | 84.3% | 81.4% |
| Decision Tree Classifier | 90.6% | 88.7% | 94.0% | 93.0% |

Overall, the decision tree classifier had the highest accuracy for both datasets, with a score of 80.0% for queue_basic and 90.6% for queue_costs. The standard logistic regression model had a slightly higher precision for queue_basic than the decision tree classifier, with 94.4% of all positively predicted values being correct. The decision tree classifier had the highest precision score for queue_costs, with a score of 88.7%. The decision tree model also had the highest recall for both queue_basic and queue_costs, with a score of 83.0% for queue_basic and 94% for queue_costs. The largest difference between the model performances for queue_basic and queue_costs was the negative predictive rate. On average, the negative predictive rate for queue_costs was about 54% higher than queue_basic. For both datasets, the decision tree classifier again performed better than the other two methods, with a negative predictive rate of 37.0% for queue_basic and 93.0% for queue_costs.

**Feature Importances—Full Datasets**

Another key takeaway from my analysis is the role of feature importances, which reveals how individual variables within the decision tree classifier significantly shape the model's predictions. I chose to focus on the feature importances from the decision tree specifically given the method having the highest performance and decision trees inherently have the most interpretability (Molnar, 2021). The power output of the facility was a key determinant for queue_basic, as shown in Figure 5.

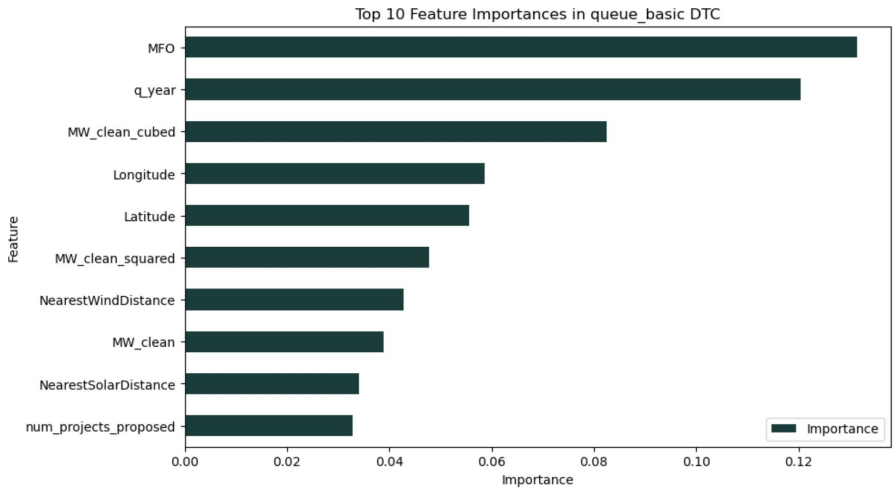

**Figure 5: Feature importances in the queue_basic decision tree classifier.** Horizontal bar chart with feature importance as the x-axis and the feature as the y-axis.

The maximum facility output (MFO) and various polynomials of the MW output make up three out of the top ten most important features in the model. Almost as important as MFO is the year the project was proposed (q_year), as Figure 5 demonstrates. Other important features for the queue_basic classifier were locational columns, such as longitude, latitude, and proximity to the solar and wind resource quality values.

In the Decision Tree classifier for queue_costs, I observed a distinct dominance of q_year, which emerged as substantially more influential than any other in the dataset. Its disproportionate weight in the model, shown in Figure 6, indicates that the model's predictions are primarily driven by this single attribute.

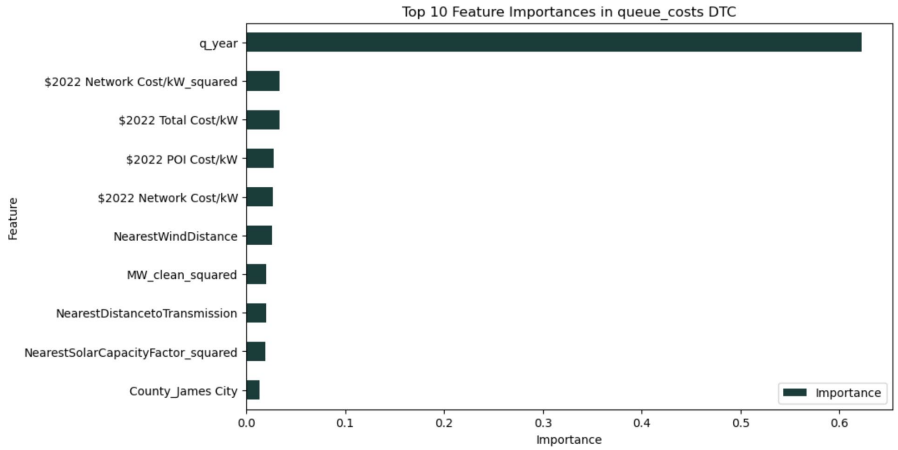


**Figure 6: Feature importances in the queue_costs decision tree classifier.** Horizontal bar chart with feature importance as the x-axis and the feature as the y-axis.

As explored in the Exploratory Data Analysis (EDA) outlined in the methods section, this outcome is expected given the drastic change in ratio between projects that execute an Interconnection Agreement or withdraw from the queue over time. Other important features include the costs of interconnection, including total cost per kW, POI cost per kW, and network cost per kW (Figure 6). Renewable resource factors also play an important role in the model.

## Model Performance—Renewable Energy Projects

An important aspect of my analysis was testing the models only on projects with a renewable energy component, given that so much of the current interconnection queue is composed of renewable energy projects. I only utilized the decision tree classifier for this task, as its performance was highest in almost all metrics on the full datasets. The performance of the model was similar for the test set with only renewable energy projects and the test set with all projects, with a slight improvement for queue_basic (Table 5).

**Table 5: Performance metrics for queue_basic and queue_costs on renewable energy projects only.** Table showing percent accuracy, precision, recall, and negative predictive rate for both queue_basic and queue_costs.

| Dataset | ACC | PRE | REC | NPR |
|---|---|---|---|---|
| queue_basic | 83.4% | 94.7% | 87.1% | 12.7% |
| queue_costs | 85.4% | 86.2% | 92.6% | 83.3% |

Overall, the performance metrics for the renewables only test set are a few percentage points higher than the performance metrics for the entire test set for queue_basic, aside from the negative predictive rate, which is notably 24% lower. The performance metrics for the renewables only test set for queue_costs are a few percentage points lower across the board. This analysis underscores the robustness of the decision tree classifier when applied specifically to renewable energy projects.

## Model Performance & Feature Importances—No Temporal Features

Given the disproportionate importance of temporal features in decision-making, I also trained and tested the decision tree classifier with optimal parameters on versions of queue_basic and queue_costs without temporal features. The features I removed were the year that the project was proposed (q_year) and the number of projects proposed in that year (num_projects_proposed). The performance metrics of the model were within a few percentage points of the initial model performance (Table 6).

**Table 6: Performance metrics for queue_basic and queue_costs with no temporal features.** Table showing percent accuracy, precision, recall, and negative predictive rate for both queue_basic and queue_costs.

| Dataset | ACC | PRE | REC | NPR |
|---|---|---|---|---|
| queue_basic | 81.7% | 94.8% | 83.7% | 36.4% |
| queue_costs | 86.5% | 85.2% | 92.9% | 88.6% |

All of the metrics for this model were within 2% of the initial decision tree classifier for queue_basic and within 5% of the initial classifier for queue_costs. All the metrics for this model were higher for queue_basic aside from the negative predictive rate, which was 0.7% lower. However, due to regularization in both models and the slight percentage differences in outcome each time the models were run, these can likely be attributed to random discrepancies. The metrics for this model were lower across the board for queue_cost, with the largest difference also being a lower negative predictive rate. The trends in the metrics are similar between the models—similar to the decision tree classifier trained on all features, the precision of the model trained on queue_basic is about 10% higher than the model trained on queue_costs, but the negative predictive rate of the model trained on queue_costs is over 50% higher than the the model trained on queue_basic.

With the removal of two of the most significant features from the initial prediction, another notable outcome was the shifting landscape of feature importances. The MW output remained the most important features in queue_basic, specifically MFO and MW_clean. The energy type and locational variables also had high importance, as shown in figure 6.

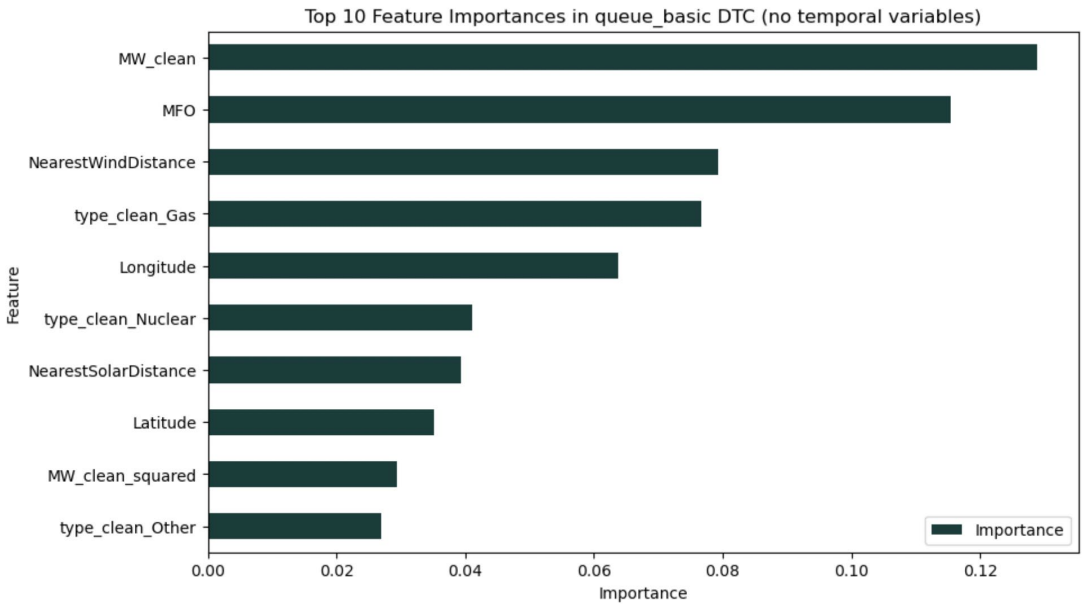


**Figure 6: Feature importances in the queue_basic decision tree classifier, no temporal variables.** Horizontal bar chart with feature importance as the x-axis and the feature as the y-axis.

With the absence of temporal variables, cost became the dominant variable, specifically network cost per kW. The next most important type was the type of project, specifically solar and battery, and the one following was the solar capacity factor of the project if the project had a solar component. Generally, features related to solar made up three out of the top ten most important features, and cost features made up two out of the top ten (Figure 7).

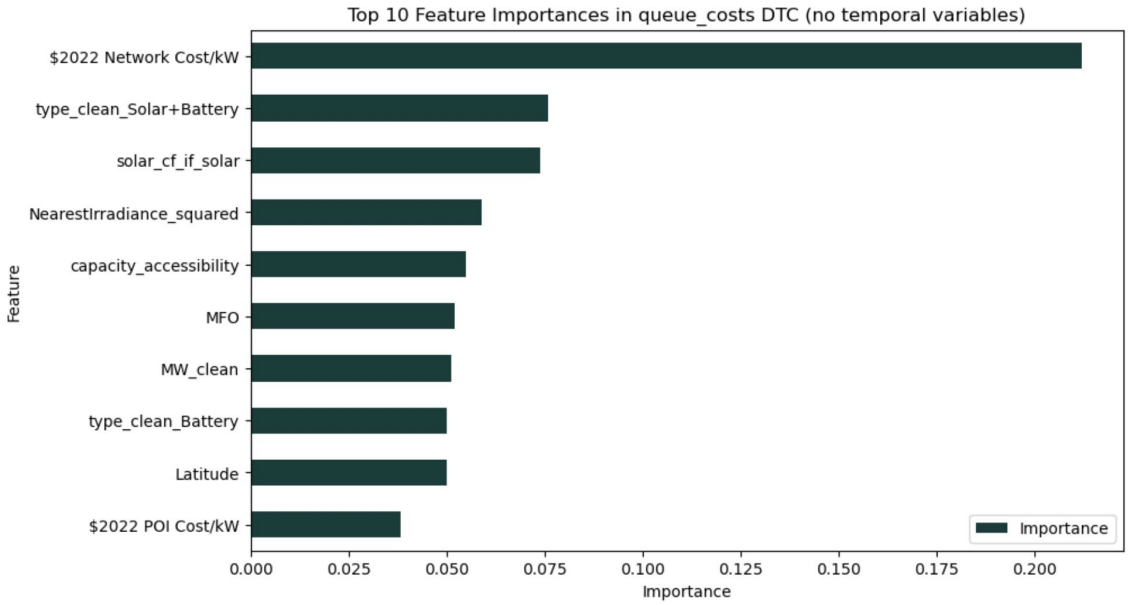


**Figure 7: Feature importances in the queue_costs decision tree classifier, no temporal variables.** Horizontal bar chart with feature importance as the x-axis and the feature as the y-axis.

Overall, this comprehensive analysis demonstrates the adaptability and consistency of the decision tree classifier across various scenarios. The analysis shows how feature importances shift when focusing solely on renewable energy projects or when removing specific types of variables, yet also demonstrates consistent performance.

## DISCUSSION

### Application of Classification Models to Interconnection Queues

Overall, my findings demonstrate the applicability of machine learning classification techniques to predicting interconnection queue outcomes. Currently, there is a lack of research exploring how to mitigate uncertainty in the interconnection in its current state, yet renewable energy developers point to interconnection as the most significant obstacle they encounter (Driscoll, 2022). While the cost of remaining in the queue is low, it creates an externality that affects all other developers. The bottleneck slows down the process of conducting studies for all projects in the queue, and many of these studies are on projects that will never reach commercial operations (Yang et al., 2023). Reducing this congestion has multiple positive impacts. It has

been shown that reducing wait time significantly increases completed capacity, specifically renewable capacity. Shortening interconnection queue wait times also reduces the generator's waiting costs such as land payments and permits as well as needing to withdraw for other reasons such as on long term contracts (Yang et al., 2023). Machine learning has proven effective in countless industries, such as financial services, healthcare, and manufacturing, at optimizing decision making and improving efficiency (Karunakaran, 2023). By predicting interconnection queue outcomes, these models can help developers make more informed decisions on whether or not to remain in the interconnection queue, potentially increasing informed withdrawals and reducing congestion.

The models I created demonstrate a marked improvement on the uncertainty of the status quo through their high performance metrics. The decision tree classifier, which performed the best, predicted the outcome with an accuracy of 80.0% for queue_basic and 90.6% for queue_costs. These metrics are much higher than the benchmark accuracy, which is 22.9% for queue_basic and 45.8% for queue_costs. Both datasets had high precisions, with 93% of the decision tree classifier's positive predictions being correct on queue_basic compared to 88% for queue_costs. These metrics indicate the models, specifically the one trained on queue_basic, are almost always correct when they predict a project will receive an interconnection agreement. The corresponding high recall values of 83% for queue_basic and 94% for queue_costs indicate that almost all of the executed projects were correctly identified by the models.

The only area where performance dipped significantly was the queue_basic classifiers' negative predictive rate (NPR), which was 37% for the decision tree classifier. This metric indicates that the queue_basic model is generating a higher level of false negatives, meaning that the model will more frequently predict a project will withdraw from the interconnection queue when it may ultimately execute an interconnection agreement. However, the NPR of queue_costs is 93%, meaning that the false positive rate is low. This disparity makes sense given that interconnection costs are a critical factor in developers' decision to withdraw from the queue, indicating the addition of these features helps identify when a project will ultimately withdraw (Yang et al., 2023). Overall, both models exhibit a high degree of accuracy in their predictions and demonstrate a proficiency in identifying projects that will successfully execute an interconnection agreement. Therefore, if a developer were to utilize these models to forecast the execution of an interconnection agreement, there would be a heightened likelihood of this

prediction being accurate. Conversely, if the developer had cost information and used the model trained on queue_costs and received the withdrawn outcome, it could incentivize them to remove their project from the queue. This could reduce congestion in the interconnection queue, increasing built capacity and diminishing resource waste among developers.

**Key Determinants of Model Outcomes**

The most important features employed for the model's predictions demonstrate significant real-world implications regarding the factors that influence a project's capacity to connect to the queue. For queue_basic, the most important features are maximum facility output (MFO) and the year that the project entered the queue. For queue_costs, the most important factor was by far the year the project entered the queue. The importance of the proposed year is logical, considering the characteristics of the datasets, especially queue_costs, and the evolving approvals in the interconnection queue in recent years. As discussed in the Methods section, the proportion of projects that execute an interconnection agreement compared to withdrawn changes dramatically over the years in queue_costs, whereas it remains more balanced in queue_basic, indicating that this shift is partially a result of the data availability. The gap in the data is driven by a decision by the LBNL to only include new generation facilities, entries that had completed a feasibility study, projects that were not superseded, and projects entering the queue before March 2021. Even with this small subset, compilation of the dataset required manual cost extraction from study pdfs, which creates a major information barrier for both researchers and prospective developers (Seel et al., 2023).

The significance of the proposed year feature also demonstrates how withdrawal rates have increased over time. Interconnection times have increased over time, and late-stage withdrawals have become much more common, which can be expensive for developers and disrupt assumptions made by other projects. In 2023 alone, the interconnection backlog grew by 30% (LBNL, 2024). Additionally, more submitted capacity has been withdrawn in the past few years, with MISO reporting that the most active interconnection customers have been removing disproportionately high levels of capacity from the queue in recent study cycles (CRA, 2023). The importance of this feature in the models is further evidence of this trend.

While the significance of the 'proposed year' feature emphasizes the increasing congestion in the interconnection queue over time, its presence in the model may not aid future developers in assessing their chances of receiving an interconnection agreement. Therefore, I also trained and tested the models on versions of queue_basic and queue_costs that did not have any temporal features. The model performance for queue_basic remained about the same, and decreased slightly to 86.5% accuracy for queue_costs, demonstrating the flexibility of the models. The most important features for queue_basic were still MFO and other MW output features, as well as location features such as longitude and latitude. The most important features for queue_costs were the network cost / kW, whether a project was solar, and solar resource quality variables. The dominance of cost as the most significant variable in queue_costs is logical, given that it represents the primary consideration for developers—renewable generators have stated that the costs and timeline of interconnection represent the largest barrier to widespread renewable adoption (Driscoll, 2022). Additionally, high costs are a key factor in generators' decision to withdraw from interconnection queues (Yang et al., 2024) Variables related to solar having a high importance is also understandable given the influx of solar into interconnection queues and onto the grid in recent years. Solar currently accounts for the largest proportion of generation capacity in the interconnection queue, and solar and storage are by far the fastest growing resources in the queues, accounting for 80% of new capacity in 2023 (Rand et al., 2024). These feature importances are in line with recent trends and demonstrate the interpretability of the model.

The importance of MFO for queue_basic and its relatively lower importance for queue_costs is logical given the nature of the datasets. Almost all of the projects that are withdrawn from queue_basic have a size between 0 and 250 MW, compared to a more even balance in queue_costs (Figure 8).

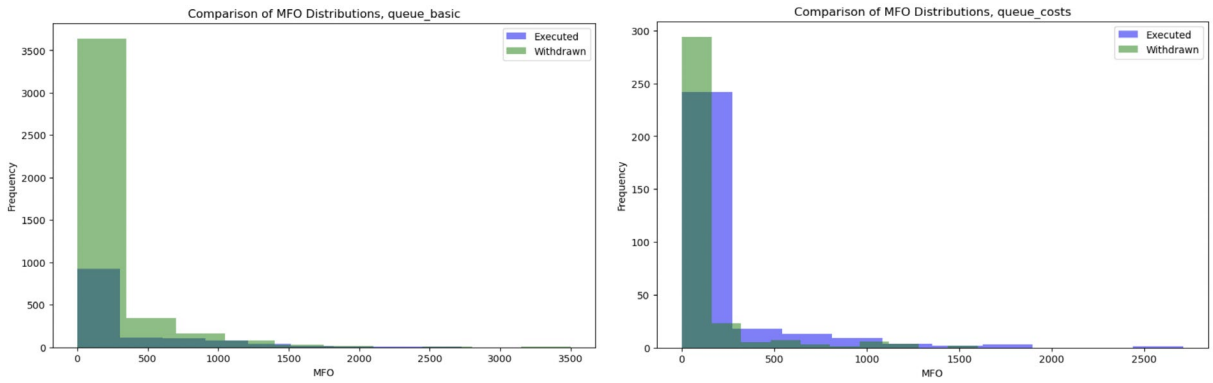


**Figure 8: Comparison of MFO distributions for executed vs withdrawn projects, queue_basic (left) and queue_costs (right)**. Overlaid bar charts with executed and withdrawn projects along the scale of MFO values. The x-axis is MFO values and the y-axis is the frequency of each project.

Many of these projects may be "placeholder" projects placed by developers in the queue to hold their place in line or to collect information on which project has the highest chance of receiving approval (DOE, 2021). Since the entries in queue_costs are further developed, likely not as many of them are placeholder projects. Placeholder projects in interconnection queues lead to several issues, including increased congestion, misallocation of resources, and higher costs for other developers due to necessary but potentially wasteful system impact studies. They can distort market signals, delaying necessary grid improvements and reducing the transparency and trust in the interconnection process. The prevalence of potential placeholder projects in the queue indicates a need for reforms that disincentivize their placement and prioritize projects based on actual feasibility.

The importance of locational factors such as latitude and longitude in queue_basic may also demonstrate the importance of locational marginal pricing (LMP) in the interconnection queue. Locational marginal pricing is a method used in electricity markets to determine the price of electricity at various locations within the electrical grid. LMP can influence the interconnection queue, as a higher LMP indicates a higher demand for electricity in a certain area and a higher return on investment for generators, which may motivate developers to pay higher interconnection costs or motivate transmission system operators to approve a project located where demand is higher (Walsh, 2023). The models can help approximate the impact of LMP on a project's probability of being approved. This insight is particularly valuable for optimizing project siting decisions and enhancing the overall efficiency of resource allocation within the energy market.

Overall, the model determinants highlight both the interpretability of the models and the significant trends evident in the interconnection queues, specifically what factors are important to project approval. By identifying temporal and locational variables such as the proposed year and locational factors, the models underscore the growing complexity and congestion in interconnection processes. These insights not only facilitate a deeper understanding of market dynamics, but also point to the need for policy reforms aimed at streamlining interconnection queues and prioritizing truly feasible projects.

**Model Performance on Renewable Energy Projects**

Currently, the vast majority of projects in the interconnection queue are renewable energy projects. There is over 1 terawatt (TW) of solar and 360 gigawatts (GW) of wind in interconnection queues around the United States, with wind, solar, and storage making up 95% of active queue capacity (Rand et al., 2024). Therefore, the performance of the model on only renewable energy projects is more relevant to the current state of the queue. When tested on only projects with a clean energy component (solar, wind, or storage), the model performed better by a few percentage points on every performance metric on queue_basic except for the negative predictive rate, which was notably lower. This outcome indicates that the model was unable to capture the negative outcomes, meaning that there were more projects that executed an interconnection agreement than the model predicted. The model trained on queue_costs performed a few percentage points lower, but still over 80% for all metrics and with a recall of 92.6%. This score demonstrates that the model correctly classified over 90% of all the positive values in the dataset. Overall, these outcomes demonstrate that the decision tree classifier also exhibits high performance when tested on only renewable energy projects, meaning that the model is relevant for renewable energy developers and the current state of interconnection queues.

**Limitations and Future Directions**

While the model performed with high performance metrics, the process of developing the model also illuminated key limitations. First, the lack of data availability remains a key issue in

the study of interconnection queues. As discussed earlier, the lack of cost of interconnection data limited the number of entries in queue_costs and thus the flexibility of the model. Additionally, the precise location of a proposed project has major implications for interconnection queue outcomes, as location determines distance to a point of interconnection, transmission congestion, and the actual renewable resource availability. For many of the projects in these datasets, I approximated location using the county's centroid, which is not an accurate measure of where the project is situated. Lack of data availability regarding the exact position of the project inhibits model accuracy. Besides constraints on the quantity of data, there are also limitations concerning data integrity. The growing phenomenon of developers entering placeholder projects into the interconnection queue has the potential to skew the data and produce inaccurate trends, as seen with the high volume of withdrawn projects producing below 250 MW. These consequences can lead to bias in the model and hinder its flexibility and accuracy on legitimate projects in the queue.

Other limitations are the constantly changing nature of policy and interconnection queues. Renewable energy targets on the state and utility level are currently featured in the model, but these goals are constantly changing and new types of renewable incentives are continually being adopted on various levels. In order for the model to remain relevant, these updates must be consistently performed, and new features may need to be added.Additionally, interconnection queues, such as the PJM queue, are undergoing significant reforms due to the escalating wait times associated with interconnection approvals. For the past two years, PJM has been working with stakeholders to change the process, and has revised its technical study process to integrate new renewable resources onto the grid more quickly and efficiently (McGlynn, 2024). While these reforms are necessary, they may change which projects are more likely to receive interconnection to the grid, potentially harming the model's accuracy.

This thesis lays the foundation for using machine learning to predict interconnection queue outcomes, but there are many potential future steps to be taken. Significantly, the datasets queue_basic and queue_costs are far from comprehensive. There are many other factors that determine whether or not a project will receive and execute an interconnection agreement. Industry professionals describe their largest barriers to renewable energy development as local ordinances and community opposition as among the most important reasons for delays and cancellations of renewable energy projects (Nilson et al., 2024). In an effort to address these

challenges, a highly meaningful inclusion would be ensuring the accuracy of the location of the proposed facilities. Once the correct lat long coordinates of a project has been determined, many other features can be added to the dataset, such as distance to a point of interconnection or transmission lines, specific town or municipal ordinances, locational marginal pricing, grid congestion, and local population. These variables could improve model accuracy and relevance to developers significantly.

Further, the datasets could be leveraged for related predictive analyses that extend beyond determining interconnection queue outcomes. Specifically, using queue_costs to develop a model that could predict the costs associated with interconnection could be very useful for developers, as cost of interconnection generally can remain relatively uncertain through multiple interconnection studies, by which point the project has already undergone development. There are also many factors which can create uncertainty regarding the cost of interconnection, as some generators leaving the queue can increase the cost for other generators. These sudden changes in cost are also generally unexpected by developers, as indicated by their behavior in the queue (Yang et al., 2024). Using machine learning to predict these costs, depending on model performance, could save developers time and money and mitigate the number of placeholder projects in interconnection queues.

Finally, while my thesis focused on PJM, there are six other RTOs that could also benefit from using machine learning to reduce uncertainty in their interconnection queues. Currently, there is over 1 TW of storage capacity in interconnection queues around the United States, primarily in CAISO and the west generally, but storage completion rates are only 11% (Rand et al., 2024). Using a model to reduce uncertainty could reduce the number of projects in queues around the US that will never be built, reducing queue congestion and saving developers millions of dollars.

## Conclusion

In conclusion, this thesis demonstrates that the application of machine learning models to predict outcomes in interconnection queues is viable for improving decision-making processes in the energy sector and reducing uncertainty in interconnection queues. Given the rapid growth and key role of renewable energy resources in the energy transition, this predictive power can be

particularly impactful for accelerating the deployment of renewable projects. Classification models such as logistic regression and decision tree classifiers offer an advancement in dealing with the inherent uncertainties of interconnection queues, providing developers with valuable insights that can lead to more informed decisions. The accuracy and reliability demonstrated by the decision tree classifier, particularly, highlight the potential of these tools to reduce queue congestion and optimize resource allocation. By accurately predicting interconnection outcomes, these models help developers avoid unnecessary costs and streamline project timelines, thus promoting renewable energy deployment.

Furthermore, while the models have shown great promise, the constantly changing nature of energy policy and the evolving landscape of interconnection processes pose continuous challenges that require ongoing adjustments and improvements to the modeling approaches. Future research should focus on expanding the datasets to cover more diverse scenarios and incorporate additional predictive variables that could affect interconnection outcomes, such as specific locational factors and policy changes. Enhancing the granularity and accuracy of the data, particularly in terms of project locations and costs, will be crucial for maintaining the relevance and accuracy of these models. Ultimately, however, while machine learning presents a way to reduce uncertainty in interconnection queues, the most impactful option is continuing to reform the interconnection process in a way that supports both developers and RTOs.

## ACKNOWLEDGEMENTS

I would like to thank my instructors, Patina Mendez and Jessica Craigg, for their invaluable guidance throughout my thesis writing process and for helping in the organization of my thoughts over the past year and a half. I am also very grateful to my peers for their support and meticulous review of my thesis, and my friends for providing moral support and overarching advice. Lastly, I would like to thank my mentor, Joe Rand, and my former professor, Duncan Callaway, whose instrumental insights helped me navigate the complexities of interconnection queues, manage vast quantities of data, and contextualize my work. You have all been integral to my success, and I am profoundly appreciative of your contributions.

## REFERENCES

Amman, D. 2023. Breaking Through the PJM Interconnection Queue Crisis. NRDC.

Benefits of Renewable Energy Use. 2017, December 20. . https://www.ucsusa.org/resources/benefits-renewable-energy-use.

Cannon, G., and P. Wiseman. 2022, July 13. The FERC (Inter)connection: PJM files for and FERC issues proposed reforms of the interconnection process. Allen & Overy.

Cantafio, R. A., and M. C. Nowak. 2021. Solving the Interconnection Problem. Texas A&M Journal of Property Law 7:526–538.

Catsaros, O. 2023. Global Low-Carbon Energy Technology Investment Surges Past $1 Trillion for the First Time. BloombergNEF.

Cifor, A., P. Denholm, E. Ela, B.-M. Hodge, and A. Reed. 2015. The policy and institutional challenges of grid integration of renewable energy in the western United States. Utilities Policy 33:34–41.

Crouch, B., Y. Dave, K. Grover, I. Gupta, M. Phan, M. Shafaie, M. Shen, and L. Weng. 2023. Principles and Techniques of Data Science. UC Berkeley.

Draxl, C., and B. Mathias-Hodge. 2016. WIND Toolkit Power Data Site Index. [object Object].

Driscoll, W. 2022, February 14. Interconnection delays and costs are the biggest barrier for utility-scale renewables, say developers. pv magazine.

Egan, D. 2015, September 17. PJM Queue Position Establishment Background and Support for Problem Statement.

Frequently Asked Questions (FAQ). 2024. . https://sepapower.org/utility-transformation-challenge/frequently-asked-questions-faq/.

Friedman, J., T. Hastie, and R. Tibshirani. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software 33.

Glazer, G., K. Siegner, C. Teplin, and S. Toth. 2022. Scaling Clean: Assessing Market Options for Clean Energy and Capacity in PJM. RMI.

Handmaker, R. S. 1989. Deregulating the Transmission of Electricity: Wheeling under P.U.R.P.A. Sections 203, 204, and 205. Washington University Law Quarterly 67:435–460.

Hodge, B.-M. 2015. Final Report on the Creation of the Wind Integration National Dataset (WIND) Toolkit and API. National Renewable Energy Laboratory, Louisville, Colorado.

Interconnection 101. 2023. . American Clean Power.

Interconnection Process Reform. 2024. . https://www.pjm.com/planning/service-requests/interconnection-process-reform.

Jones, L. E., editor. 2017. Renewable energy integration: practical management of variability, uncertainty, and flexibility in power grids. Second edition. Academic Press is an imprint of Elsevier, London, United Kingdom ; San Diego, CA, United States.

Karabiber, F. 2024. Gini Impurity. https://www.learndatasci.com/glossary/gini-impurity/.

Karunakaran, K. 2023, May 18. The Role of Machine Learning in Automating Decision-Making Processes. LatentView Analytics.

Khan, K., and C. W. Su. 2022. Does policy uncertainty threaten renewable energy? Evidence from G7 countries. Environmental Science and Pollution Research 29:34813–34829.

Khempila, A., and V. Boonjing. 2010. Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients. 2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM):193–198.

Lenor, T. 2023. US Interconnection Queues Analysis 2023. S&P Global.

Liu, L., F. Bai, C. Su, C. Ma, R. Yan, H. Li, Q. Sun, and R. Wennersten. 2022. Forecasting the occurrence of extreme electricity prices using a multivariate logistic regression model. Energy 247:123417.

Lopez, A. 2021. Solar Supply Curves. https://www.nrel.gov/gis/solar-supply-curves.html.

Mays, J. 2023. Generator Interconnection, Network Expansion, and Energy Transition. IEEE Transactions on Energy Markets, Policy and Regulation:1–10.

McGlynn, P. 2024, April 23. Interconnection Reform Is Working, but Will New Generation Actually Get Built? https://insidelines.pjm.com/interconnection-reform-is-working-but-will-new-generation-actually-get-built/.

MISO Interconnection Queue: M2, M3 and M4 security deposits and return procedures. 2023, August 26.

Molnar, C. 2023. Decision Tree. Page Interpretable Machine Learning.

Moving Through the Interconnection Queue: How a Project Gets Built—or Doesn't. 2023. Advanced Energy United.

Nilson, R., B. Hoen, and J. Rand. 2024. Survey of Utility-Scale Wind and Solar Developers Report. Lawrence Berkeley National Laboratory.

Nudell, T. R., A. M. Annaswamy, J. Lian, K. Kalsi, and D. D'Achiardi. 2018. Electricity Markets in the United States: A Brief History, Current Operations, and Trends. Switzerland.

PJM - Who We Are. 2023. https://www.pjm.com/about-pjm/who-we-are.aspx.

Porter, K. 2002. The Implications of Regional Transmission Organization Design for Renewable Energy Technologies. National Renewable Energy Laboratory, Colorado.

Proposed Power Plants Point to a Clean-Energy Future. 2021. . Department of Energy, Office of Energy Efficiency & Renewable Energy.

Rand, J., N. Manderlink, W. Gorman, R. H. Wiser, J. Seel, J. M. Kemp, S. Jeong, and F. Kahrl. 2024. Queued Up: Characteristics of Power Plants Seeking Transmission Interconnection. Lawrence Berkeley National Laboratory.

Rudell, T., A. Annaswamy, J. Lian, K. Kalsi, and D. D'Achiardi. 2018. Electricity Markets in the United States: A Brief History, Current Operations, and Trends. Pages 3–27 Smart Grid Control. Springer Berlin Heidelberg, New York, NY.

scipy.spatial.cKDTree. 2024. . https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.cKDTree.html.

Seel, J., J. Rand, W. Gorman, D. Millstein, R. Wiser, W. Cotton, K. Fisher, O. Kuykendall, A. Weissfeld, and K. Porter. 2023. Interconnection Cost Analysis in the PJM Territory. Lawrence Berkeley National Laboratory.

Sengupta, M., A. Habte, A. Lopez, and Y. Xie. 2017, December 12. The National Solar Radiation Data Base (NSRDB).

Shorabeh, S. N., N. N. Samany, F. Minaei, H. K. Firozjaei, M. Homaee, and A. D. Boloorani. 2022. A decision model based on decision tree and particle swarm optimization algorithms to identify optimal locations for solar power plants construction in Iran. Renewable Energy 187:56–67.

Srinivasan, A., R. Wu, P. Heer, and G. Sansavini. 2023. Impact of forecast uncertainty and electricity markets on the flexibility provision and economic performance of highly-decarbonized multi-energy systems. Applied Energy 338:120825.

Ulkhaq, M. M., A. K. Widodo, M. F. A. Yulianto, Widhiyaningrum, A. Mustikasari, and P. Y. Akshinta. 2018. A logistic regression approach to model the willingness of consumers to adopt renewable energy sources. IOP Conference Series: Earth and Environmental Science 127:012007.

Utility Carbon-Reduction TrackerTM. 2024. . https://sepapower.org/utility-transformation-challenge/utility-carbon-reduction-tracker/.

Walsh, B. 2023, November 29. Understanding LMP & The Interconnection Queue. https://www.landgate.com/news/understanding-lmp-the-interconnection-queue.

Wilson, J. D., R. Seide, R. Gramlich, and J. M. Hagerty. 2024. Generator Interconnection Scorecard. Advanced Energy United.

Zhou, V. 2022, September 16. A Simple Explanation of Information Gain and Entropy. https://victorzhou.com/blog/information-gain/.

## APPENDIX

The code is available publicly in this git repository. The datasets, as well as a .ipynb file and a webPDF of the code, can be found in this Google Drive folder.