Andrew Robbertz
December 12, 2021

Identifying Strategic Locations
for Renewable Energy Investment

1. Introduction

Switching from non-renewable energy sources to sustainable and renewable energy sources is critical for reducing the impact of climate change in the decades and centuries to come. However, there are many complications in implementing renewable energy for large populations. The geography, weather, and development of an area must be able to support one of several forms of renewable energy including solar, hydroelectric, wind, or geothermal energy. The infrastructure of the area must be able to support the transmission of power to distribution locations and eventually to homes, businesses, and community buildings. The solution and equipment require skilled workers to operate and maintain unless automated. Finally, the demand for renewable energy in the area must match the capacity and willingness to invest in such solutions.

Each solution for renewable energy has its own requirements for the geography and weather of the area where it is to be implemented. For example, solar power implementations may be best suited in areas that have long periods in the sun, not obstructed by cloud cover, rain, snow, or tree coverage. Hydroelectric implementations require immediate proximity to a flowing body of water. Wind-powered solutions may be best suited in areas with high continuous wind speeds and have no relation to other weather patterns. Additionally, each solution has its own requirements for the infrastructure of the area where it is to be implemented. High power transmission lines must be available or created to carry the energy, with minimal loss, to nearby cities or larger distribution centers.

It is often difficult to quantify the positive impacts of implementing renewable energy solutions. Their implementation is often more expensive in up-front costs in comparison to non-renewable counterparts. The availability of energy through these solutions can be volatile based on the seasons, leading to deficiencies in some periods and potentially excess in other periods. The cost of maintenance can be very high depending on the solution being used, requiring recurring investments in the solution being used. It is also unclear how demand for energy in each area may change in the future based on changes in population density and industry.

Many factors contribute to a location being suitable for renewable energy in general. There must be sufficient demand for energy so that minimal excess and loss are created. The area must have sufficient capital to invest in costly solutions. Finally, there must be a culture and motive among the population of the location to invest further, build, and maintain the complex solutions for renewable energy.

Based on all these factors, it is difficult to select areas of the country and world where investing in renewable energy will have the greatest impact and return on investment. Therefore, it is critical to develop models for identifying locations that are best suited for

investment in renewable energy. This paper proposes a system for identifying locations that are best suited for investments in renewable energy based on a multi-factor analysis.

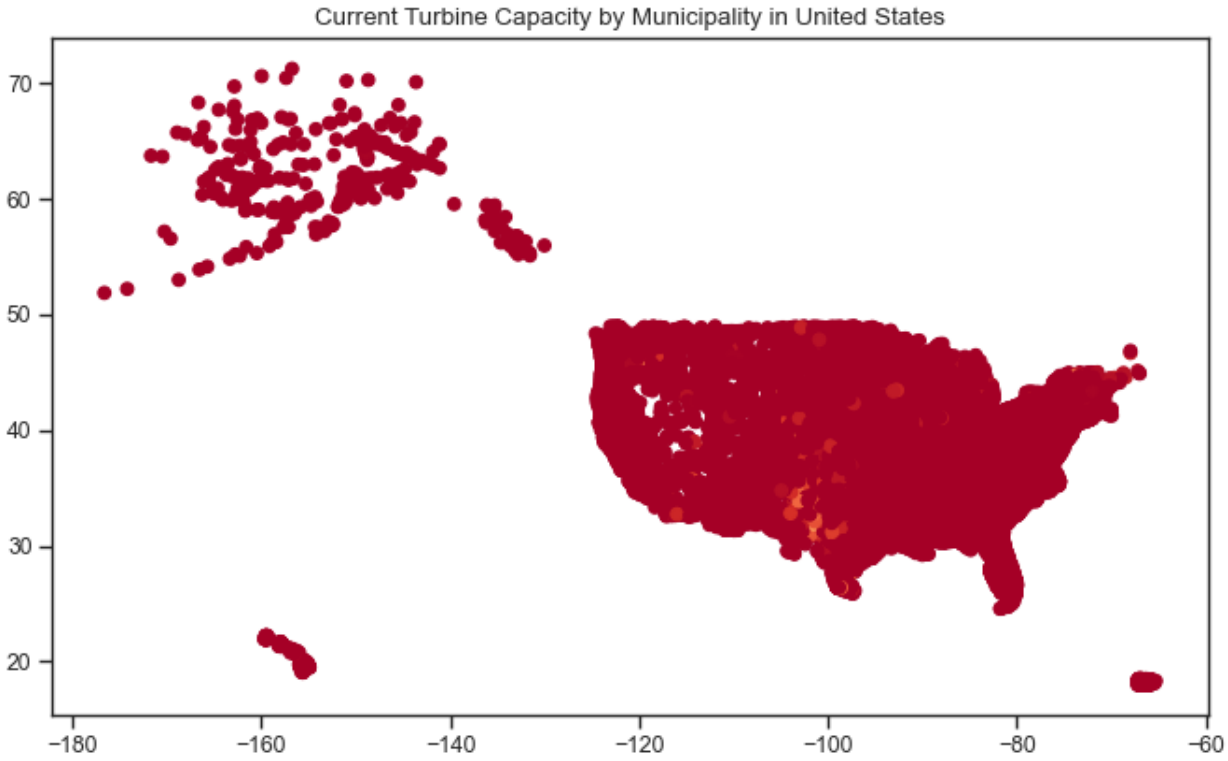
2. Decomposition

Identifying areas that will benefit most from investment in renewable energy involves many factors including the geography of the area, weather patterns, power transmission infrastructure, sufficient capital in the community to invest, and willingness to support and sustain the investment in the future. This activity requires domain experts from various fields including engineering and urban planning and can be time and resource intensive. Experts must work together to analyze the relevant data, create predictive models for one of many solutions, and generate reports for investors and policymakers.

An artificial intelligence solution to make recommendations for locations to invest in renewable energy would reduce the reliance on and workload of domain experts needed. An artificial intelligence solution could be given direct, automated access to the same quality data sources as domain experts in their relevant fields. The system or system of systems could be engineered to generate recommendations, reports, and analysis from each specific area of interest, as well as provide final predictions on the outcome of implementing renewable energy in specific areas. This would free the time and resources of domain exports to simply review the analysis of the AI system while being able to focus on other projects.

Additionally, feedback could be obtained in real-time for investors interested in creating new solutions for renewable energy. Preliminary predictions could be created on-demand to understand at high levels the effectiveness of their investment. This would allow investors to look at more areas in a shorter period of time before asking domain experts for further analysis and suggestions.

Finally, an artificial intelligence solution may be able to identify other factors, not previously considered by investors or domain experts, that lead to the effectiveness of a renewable energy implementation. The AI solution would be able to handle substantially more data than any individual domain expert. Additional data, perhaps suspected to have little to no impact, can be used to train such a model. This data may turn out to be more relevant than previously thought, identifying new key elements and considerations for how renewable energy is invested in. New insights on the factors that lead to successful investments can give further context to domain exports and investors that can be used on future projects and can be tailored to their individual expertise.



A fully integrated solution to encapsulate all factors and make recommendations for all implementations of renewable energy would be a significant undertaking. Therefore, it is best to decompose the larger system, of identifying locations that are best suited for investments in renewable energy, into multiple smaller systems, for identifying locations there a best suited for investments in a particular solution or implementation of renewable energy. For the simplicity of design and the ability to create rapid minimum viable products, we focus the remainder of this paper on an artificial intelligence system to identify locations that will best be suited for investment in wind energy. Implementations of renewable wind energy have the fewest geographical, geological, and environmental constraints making it a good candidate for a proof-of-concept system and application. Included with this paper is a demonstrative system showing the ability of artificial intelligence algorithms to draw data from several enterprise-level sources and make predictions on the output of a wind energy implementation in a particular area. Although an ideal system would not be limited to any area, the demonstration is limited to the United States and Puerto Rico based on the availability of data, to reduce processing time, and to use a familiar location as a case study to be drawn upon for future work.

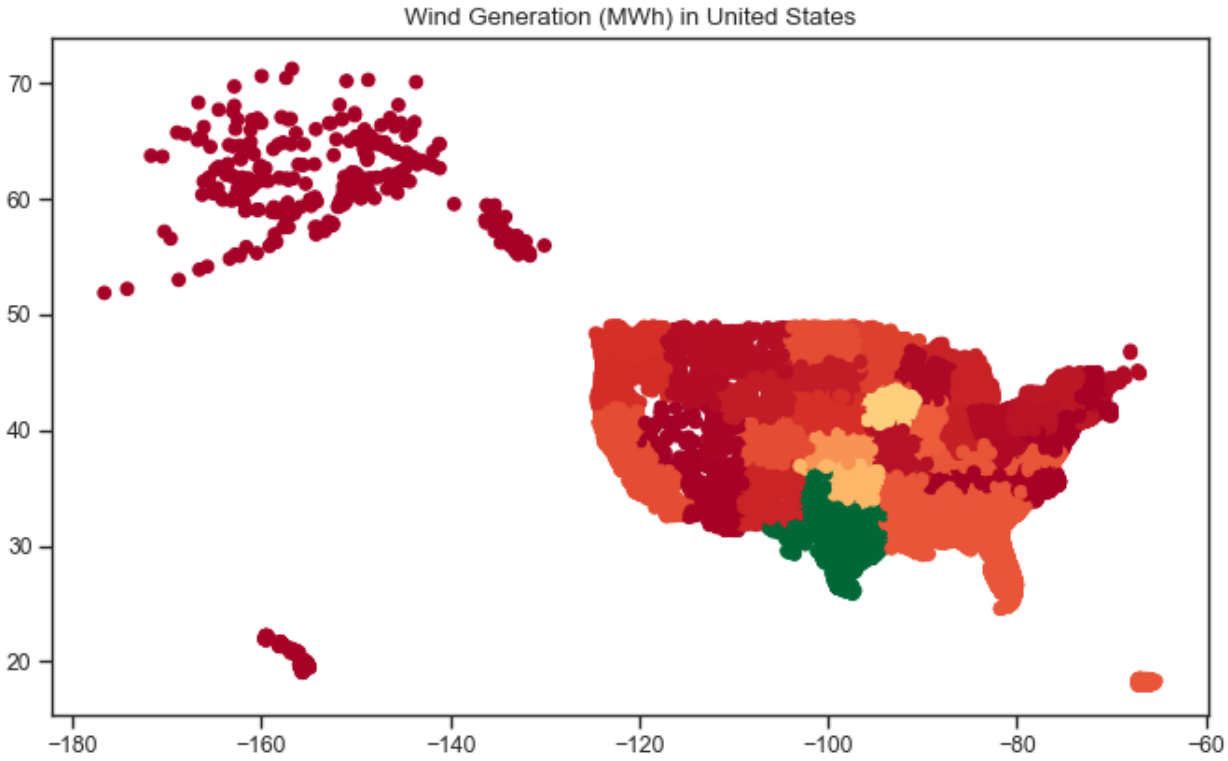
3. Domain Expertise

Domain expertise plays a critical role in the current identification of areas for investment in renewable energy. It requires input from multiple fields including engineering, electric power industry, urban planning, policymakers, environmental conservation, and finance. The same is true for the design and implementation of an artificial intelligence solution to achieve the same goal. To predict new areas that will benefit from investment in renewable energy we must first understand how to measure the effectiveness and return on investment of existing solutions.

The success of the current solution may be determined based on many factors including the amount of power produced on average year-over-year, the amount of energy lost in transmission or storage, the recurring costs of maintenance, the impact on the local environment, and animal habitats, and the cultural impact on the population living in the area. This determination must be used to create a ground-truth bases upon how to predict the effectiveness of future projects. By consulting with domain experts to obtain a ground truth for how to measure the success of a renewable energy project, we can further identify and decompose what data is necessary as an input for the system.

In addition to creating a methodology for establishing a ground truth, domain experts are critical for identifying and evaluating the datasets used for the system application. Based on their prior experience and knowledge they can identify data sources that are reliable and can be scaled to global applications. Additionally, they can assist in identifying what data points from each source are most relevant. This is critical to ensure all needed and relevant data is consumed, but little extraneous data is used to reduce the dimensionality of the data model and to improve ease of training and performance of and AI models used. Domain experts can help identify any latent features that must be generated in support of the system. By combining one or more data points in a complex feature, the system can both reduce the dimensionality of inputs while maintaining high levels of information gain from all features that are used as an input.

If any data sources contain partial, outlier, sensitive, or void data, domain expertise is required to understand how to handle such data during system processing. There are multiple strategies to handle each type of anomaly, however, the best strategy depends on the type of data. Domain experts would have this knowledge from previously analyzing similar data and their ability to draw insight on how the mitigations strategies may affect the performance of the system.



For the purposes of the system demo associated with this paper, the ground truth was selected to be a simple metric without consultation from domain experts. The ground truth metric to measure the success of an area in implementing wind energy was simply the power out generated from the wind turbines or other implementations employed. As described previously in this section, this metric alone does not encapsulate the full scope of the success of the project. However, it does enable a proof-of-concept model to predict power output as an indicator of a more complex measure of success. Similarly, data and feature engineering were conducted with little domain expertise, but to enable a proof-of-concept system using commonly accessible data and simple feature extraction based on the raw data. Finally, a standard method for handling anomalous data was selected including setting anomalous values to the mean value of similarly grouped data. This was chosen as there were relatively few anomalous entries in the dataset and the impact of this strategy did not critically misrepresent the data.

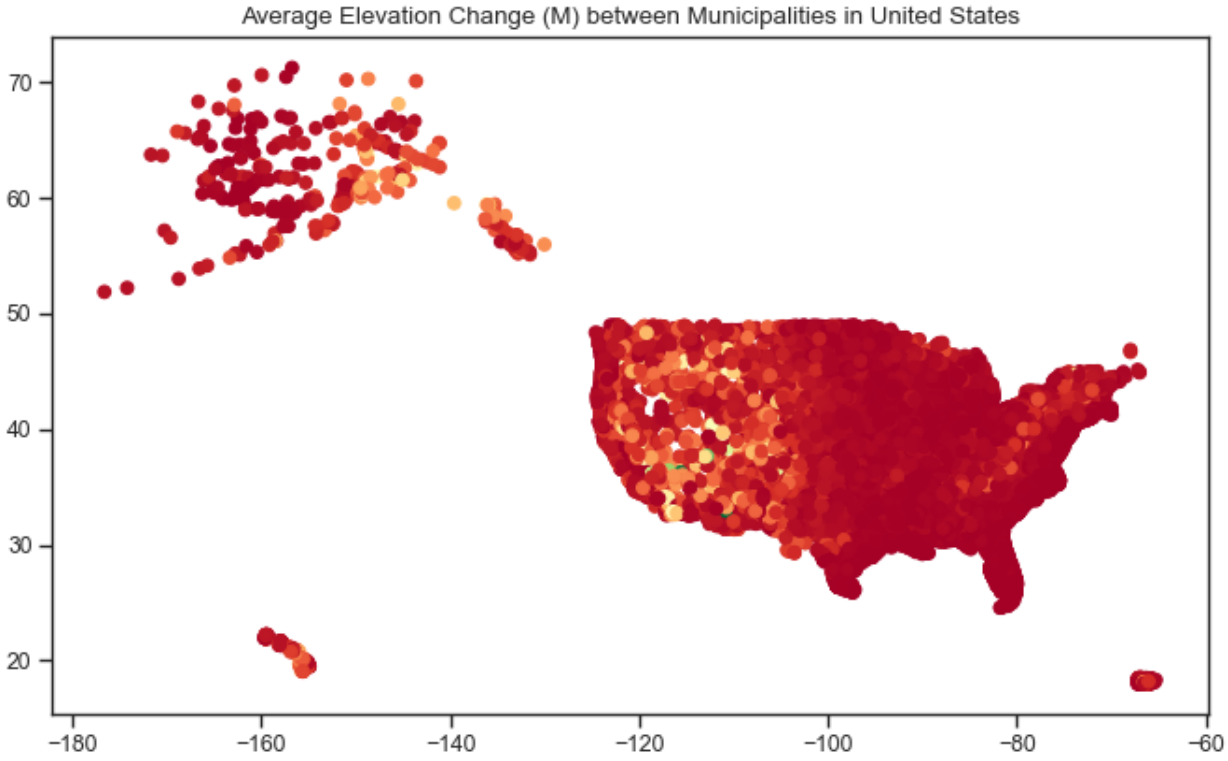
4. Data

There are a significant number of data sources from which information can be extracted to generate metrics to power our models. The forms of the data and the weight of the metrics generated to support the system's decisions and predictions should be guided by results from interviewing and working alongside domain experts. For the system demo, data was extracted from six sources:

1. Location Data: Including ~28,000 municipalities in the United States, their population, density, and neighboring municipalities
2. Wind energy production: including the total energy generated from wind power stations for each state in the United States

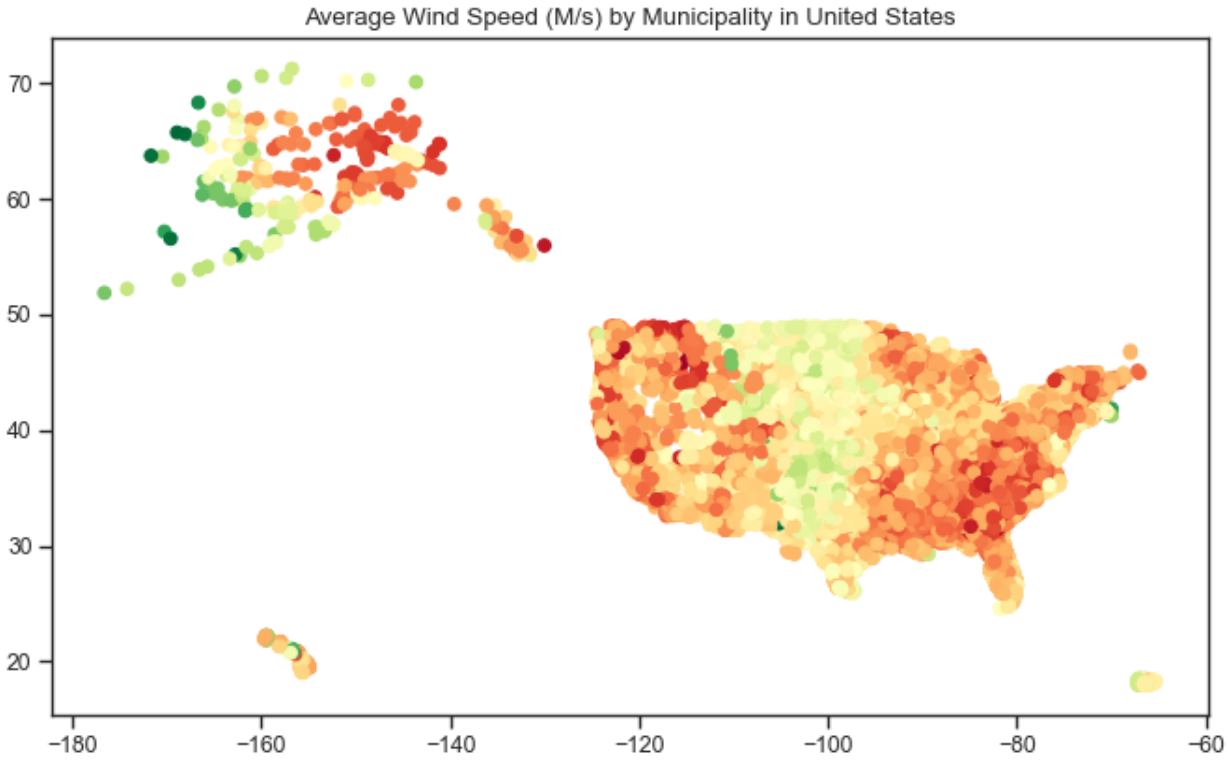
3. Turbine data: including the location, model, output, and maintenance records of ~1,100 wind turbines across the United States
4. Geological data: including the elevation for each municipality identified and average elevation change between its neighbors
5. Weather Data: including temperature, precipitation, wind, and sunlight for each of the municipalities identified based on local weather stations.
6. Income Data: including the mean House Price Index measured by the Federal Housing Finance Administration and the average change of the index between 1975 and 2020

Once the data was collected from these sources, simple derived features were generated to support the model. Although a limited number of features were generated, the model evaluation showed strong performance showing that the measures used were strongly correlated with the label chosen.



In a greater system, more relevant data should be collected based on the suggestions from domain experts. There is likely more relevant and fine-grained energy production data available. From my research, only energy production data was available at the state level, although it's likely this data can be identified at the municipality level to be used for more finely tuned predictions. Not all values from the wind turbine data were used within the data model. Additional values from the dataset, including service records and history for specific turbine models, could be used to better predict outcomes of a specific area and implementation. Additional geological data would improve the data model including tree coverage and density. For other implementations of renewable energy, relevant geological data would be needed for hydroelectric or geothermal predictions. These sources could be identified and analyzed in partnership with domain experts to determine the best sources features, and weights to be used for each. The accuracy of weather data was limited to the last year, 2021, which may not show

all trends relevant for the long-term prediction of weather patterns. Domain experts may use predictive weather models in their analysis to understand how the weather patterns of an area may best suit one or more implementations. Finally, additional income and community data are necessary to understand whether a municipality has the capital and ability to install and sustain a renewable energy implementation in the long term.



Finally, with the help of domain experts from other countries, this data should be scalable to a worldwide model. Although some data sources may not be available for worldwide modeling to the same granularity as in the United States, the system may be able to be deployed such that it can handle multiple levels of granularity concurrently depending on the data available.

5. Design

There are many models applicable for our system to identify areas that would benefit from investment in renewable energy. Unsupervised, supervised, and reinforcement learning could all be used to train a machine learning model for making such predictions.

For the system demo created, we tested two implementations of unsupervised learning including K-Means clustering and DBSCAN to identify how well the locations could be clustered into groups. The clustering methods used simple point-wise distances to identify how similar locations and data points were in the n-dimensional vector space created. The K–Means models were evaluated with varying numbers of clusters and distance thresholds to identify how well the cluster labels correlated with the target variable.

Three implementations of simple supervised learning were used to predict the target variable of the amount of power that could be generated from a given location. A kernel-based

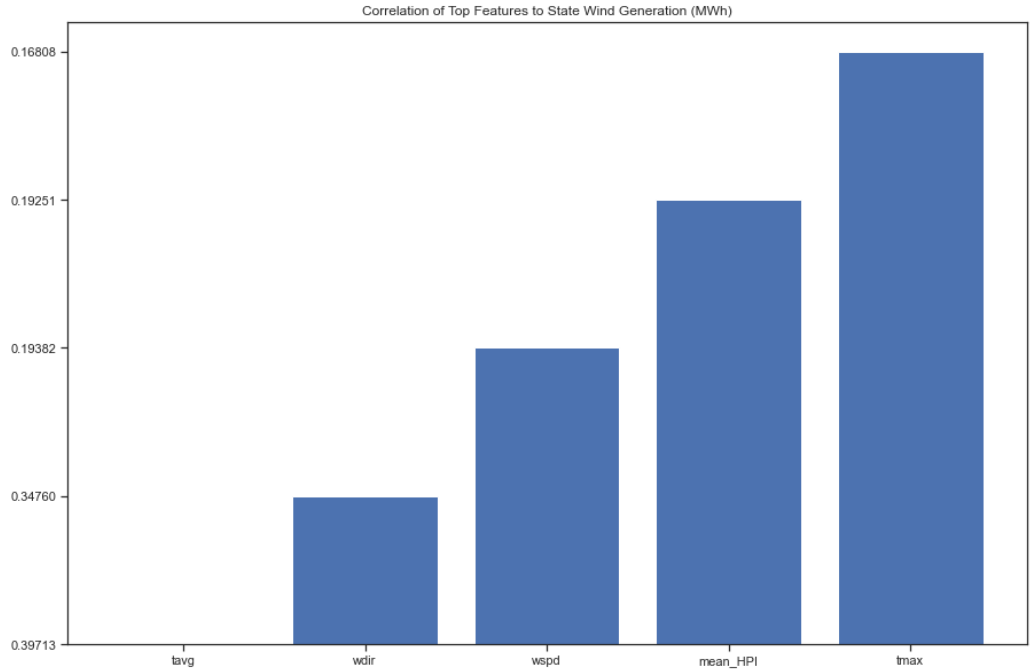
support vector regression model, a linear regression model, and a stochastic gradient descent regression model were all tested using 5-fold validation on the dataset of locations across the United States.

Finally, an Azure AutoML experiment was run using 47 unique models to further evaluate the feasibility and quality of the system demo data model. The system demo data was uploaded to Azure blob storage and identical columns were used for the experiment as the manually trained supervised learning regression models.

The default hyperparameter values were used for each of the manually trained models as a part of the system demo, however additional parameter tuning is possible to sure the models are performing as effectively as possible. One method for model and hyperparameter tuning to be used in the system should be selective sampling. Since there are a wide array of models that could be used for this task, many models and hyperparameter settings should be tested on a subset of the global data used for the greater system.

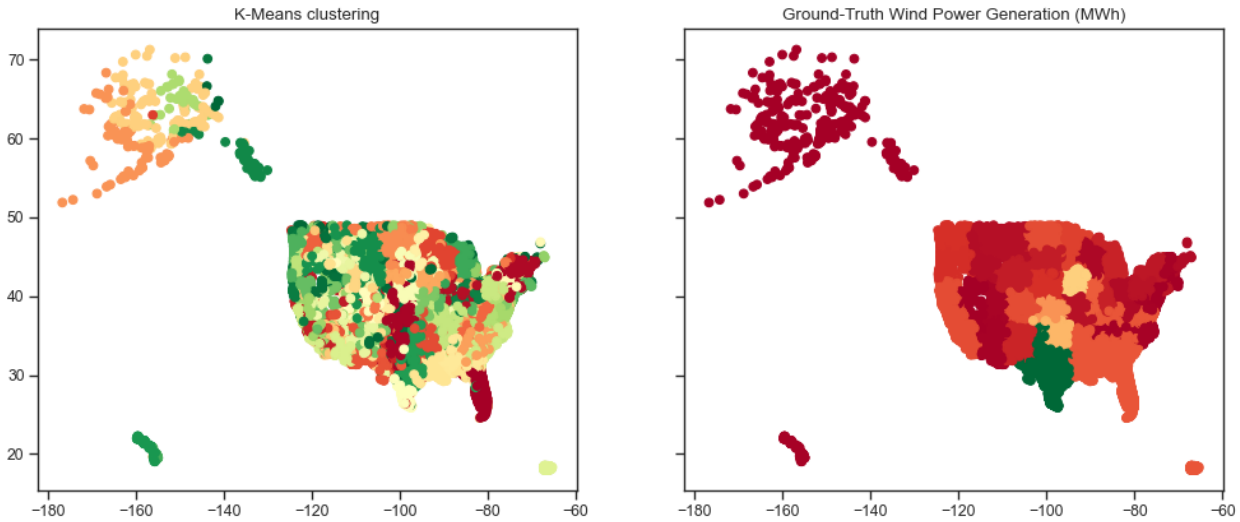
6. Diagnosis

The system demo data model was evaluated by calculating the absolute correlation between each feature and the ground truth label of power generated in MWh. The figure below shows the correlation of the top features.



The machine learning models were evaluated by calculating the direct absolute correlation between unsupervised and supervised predictions with the ground truth values for all locations. The data model did not perform well in unsupervised clustering with either K-Means clustering or DBSCAN. K-Means clustering labels showed a slightly higher correlation with the ground truth labels when compared with DBSCAN. This was slightly unexpected since the latter is often the better clustering tool. It is likely however that better values could have been chosen for the

DBSCAN hyperparameters leading to improved performance compared with the other clustering methods.

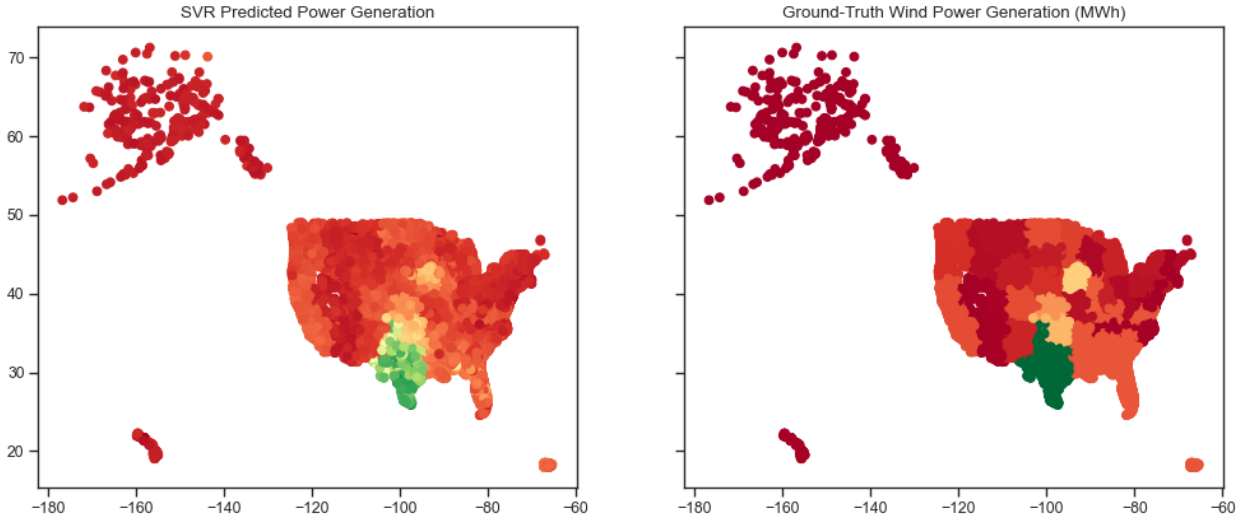


More significantly, the results from each experiment show that distance metrics alone are not capable of separating the features of the data model to accurately correlate with the ground truth. From the results of the supervised learning experiments, we can see the weighted features within the data model performed much better for predicting the ground truth. The table below shows the combined results from the manually trained models as well as the Azure AutoML training. Only the top 10 models from the Azure Auto ML experiment are shown in the table below from the total 47 models that were tested.

Algorithm Name	Normalized Root Mean Squared Error	Training Method
VotingEnsemble	0.00132	Azure AutoML
MaxAbsScaler, LightGBM	0.00134	Azure AutoML
StackEnsemble	0.00251	Azure AutoML
StandardScalerWrapper, LightGBM	0.00604	Azure AutoML
RobustScaler, LightGBM	0.00851	Azure AutoML
MinMaxScaler, RandomForest	0.01017	Azure AutoML
MinMaxScaler, LightGBM	0.01178	Azure AutoML
MaxAbsScaler, XGBoostRegressor	0.01265	Azure AutoML
MinMaxScaler, RandomForest	0.01672	Azure AutoML
StandardScalerWrapper,	0.01721	Azure AutoML

RandomForest		
Support Vector Regression	0.265	Manual
Linear Regression	0.792	Manual
SGD Regression	0.796	Manual

The table shows that the Azure AutoML regression models significantly outperformed the manually trained models using the same evaluation methods. The AutoML models, primarily ensemble models reported lower regression error for predicting the ground truth in the system demo. Further, the results from the AutoML training, as well as the best manually trained model, show that the supervised learning method can be effectively used to predict the ground truth variable based on the data model created.



There is some bias introduced in the system demo that must be resolved before a stronger system is created. The ground truth variable chosen within the system demo's design is not an accurate representation of the outcomes of implementing renewable energy in each location. The demo used simply the expected power generated in MWh, however, this does not consider the financial, environmental, communal benefits or drawbacks from the solution. Another influence in the bias of the system demo is missing data from the various datasets selected. Data cleaning and pre-processing were required to normalize the data to be used within the system demo. More complete and cohesive datasets must be identified alongside domain experts to reduce the impact of this bias within a greater system.

There are a few ethical concerns in the proposed system design and implementation of the system demo. There are few person-specific ethics concerns in this area, however, there are some community-centered ethics concerns that can be addressed. The design does not need to obfuscate or secure highly sensitive data. All the data used by the design is publicly available, either through government organizations or through accredited institutions. The main area of ethical concern is property rights. The design may suggest areas to implement renewable energy where there is high population density, where property may need to be bought or obtained from community members. Additionally, research into the community impacts of the implementation would be necessary to ensure there are no long-term adverse

effects on the population. There may also be ethical concerns in the implementation based on the environment where it would be used. The analysis would have to be done to ensure there is no long-term impact on the wildlife, air or water quality, or other ecological systems.

7. Deployment

The deployment of the system would benefit most from implementing both client-side and server-side applications. A client-side application would allow investors or researchers to input desired locations as well as any other relevant inputs. The client-side application would then send a request to the server-side application to make relevant predictions based on the data provided by the client, aggregated with relevant data stored in the server-side application or coming from live external APIs. Finally, the client-side application would receive a response with predictions for the success of the proposed implementation of wind or other renewable energy in the given location.

The server-side application would contain the fully trained machine learning model as well as API connections to each of the relevant data sources. All functionality to clean, process, and extract features must be contained in the server-side application to make on-demand predictions for the feasibility and success of implementing renewable energy in each location.

Individual deployment pipelines can be created for each aspect of the system deployment. The client-side application, the machine learning model, and each API processor can be architected to be deployed independently to enable the best decoupling of the system components. Integration tests can be executed upon each deployment to ensure the API interface for each of the system components does not change during successive deployments.

The system can be deployed entirely through Azure compute platforms. Azure provides scalable clusters for client-side web applications as well as scalable compute clusters that can host the server-side models as well as any API drivers and pre-processing needed to reach the remote databases.

8. Conclusion

The goal of the proposed system is to enable the real-time prediction of the outcomes and success of implementing renewable wind energy in each location based on a muti-factored analysis of relevant data. The system accompanying system demo shows that the required data sources are available to create a cohesive data model for solving the problem at hand. We've further identified data not included in the data model for the demo but would likely benefit the data model for future system development. Ground truth was established for the demo as the predicted power generated from a given location. Multiple unsupervised and supervised machine learning models were used to understand the quality of the data model used in the system demo. Results showed that the data model created, and rudimentary machine learning models were capable of accurately predicting the ground truth variable. New areas for improvements to improve the quality of the data model and reduce bias contained in the data were identified to be implemented in a future system based on assistance from domain experts.

Finally, a deployment strategy was created to minimize the coupling of the system and enable rapid deployments and integration of the system for end-users.

References

[1]"US Cities Database | Simplemaps.com", *Simplemaps.com*, 2021. [Online]. Available: https://simplemaps.com/data/us-cities. [Accessed: 12- Dec- 2021].

[2]B. Hoen, J. Diffendorfer, J. Rand, L. Kramer, C. Garrity and H. Hunt, "United States Wind Turbine Database (ver. 4.2, November 2021)", *U.S. Geological Survey*, 2018. Available: https://doi.org/10.5066/F7TX3DN0. [Accessed 12 December 2021].

[3]"Detailed State Data", *Eia.gov*, 2021. [Online]. Available: https://www.eia.gov/electricity/data/state/. [Accessed: 12- Dec- 2021].

[4]"Open Topo Data", *Opentopodata.org*, 2021. [Online]. Available: https://www.opentopodata.org/. [Accessed: 12- Dec- 2021].

[5]"Meteostat Developers", *Dev.meteostat.net*, 2021. [Online]. Available: https://dev.meteostat.net/. [Accessed: 12- Dec- 2021].

[6]"House Price Index Datasets | Federal Housing Finance Agency", *Fhfa.gov*, 2021. [Online]. Available: https://www.fhfa.gov/DataTools/Downloads/Pages/House-Price-Index-Datasets.aspx#qpo. [Accessed: 12- Dec- 2021].