# Photovoltaic Rooftop Installations: A Predictive Study of Market Demand

Henry Odom, Jameelah M. Young, Grace Fu, and James Forman

*Georgetown University, School of Continuing Studies, Professional Certificate in Data Science*

***Abstract*—An estimation of the demand function for photovoltaic rooftop installations in the United States produced by wrangling and munging datasets that either measure or proxy the key determinants of demand. Based on economic theory of demand, key determinants have been identified and evaluated using machine learning techniques, and operationalized as a web-based application programming interface.**

***Index Terms*—Photovoltaic, solar, PV, residential, clean energy, renewable energy sources, market forecast**

## I. Introduction

While less than 1% of total 2015 energy production in the United States was attributable to solar energy, the pace of technological advancement and promise of a clean, inexhaustible power source have made it one of the best hopes for sustaining the US's high living standards and avoiding the potentially disastrous consequences of the carbon-based economy. However, the infrastructure investments and supporting policies necessary for managing the intermittent power source and enabling local utilities to handle the decentralized nature of grid-connected PV rooftop systems are not insignificant. Accurately forecasting the number of future PV rooftop installations is critical to the timing and coordination of these policies and investments. Also, businesses interested in supporting PV rooftop system installations also desire to time their entry into new markets to be in step with potential demand and ahead of their rivals in meeting that demand. Rapid technological developments and the associated fall in PV rooftop installation costs means that the solar industry may be approaching a potential tipping point in several markets that could catch business and government policy-makers flat-footed without the right forecasting tools. Our capstone project offers policy-makers, business, and other interested parties an interactive tool for inputting simple variables, such as price, county, and/or year, and making predictions about the future demand for PV rooftop systems.

## II. Data Sources and Ingestion

Using the economic theory of demand, the team drafted a statistical abstract detailing data sources that were believed to measure the key determinants of demand for PV rooftop systems. According the general theory of demand, the key determinants of demand for any product are:

- Product price
- Income
- Income distribution
- Availability / price of complementary goods
- Price of substitute goods
- Consumer preferences
- Government policy

Utilizing the general economic determinants of demand as a guide, the team identified, ingested, and wrangled data sources believed to be accurate measures of demand determinants relevant to residential PV rooftop systems. Those data sources included:

- PV rooftop installation data collected by the National Renewable Energy Laboratory (NREL)
- PV rooftop installation data collected and cleaned by the Lawrence Berkley National Laboratory (LBNL)
- County-level income and housing stock data collected in the American Community Survey by the US Census Bureau
- Average state electric utility prices for 2015, collected by the US Energy Information Agency
- County-level Presidential General Election Results – 2012
- A-F grades of states' net metering policies and interconnection regulations / procedures produced by the Interstate Renewable Energy Council (IREC)
- 2015 National Counties Gazetteer File with latitude data by county
- County-level solar insolation data collected in the North America Land Data Assimilation System (NLDAS) Daily Sunlight (KJ/m²) (1979-2011) by NASA

Henry Odom, Jameelah Young, Grace Fu and James Forman are participants of the Professional Certificate in Data Science Program at Georgetown University, Washington, DC 20057S

The data store for the Capstone project was a PostgreSQL database hosted on Amazon Web Services (AWS) as a RDS instance. Most of the datasets detailed above were compact CSV files that could be and were copied directly into properly pre-formatted tables for WORM storage without need for Python or other scripting. However, the Tracking the Sun Public Data File and Full Open PV Dataset, due to their large sizes, had to be ingested into the AWS PostgreSQL database via Python scripting from downloaded csv's on team members' computers. To solve the data formatting compatibility issues that arose during multiple attempts to ingest these datasets, the Python scripts converted all data into strings for ingestion which the team then converted into the appropriate numeric / date formats using PostgreSQL functionality. conference page limits.

## III. Data Wrangling

After successfully ingesting the datasets above, the team had to aggregate, transform, and join the desired features from the various ingested tables into a machine learning ready dataset. Detailed below are the steps taken to create the initial machine learning dataset:

1. Instances: The team decided to use quarterly residential PV rooftop installation counts by county as the instances for its predictive model and made its first priority to create a SQL aggregate view of installation counts by county, year, quarter. Most installation records in the Tracking the Sun Public Data File had county location data that the team linked to FIPS (Federal Information Processing Standard) geography codes for joining and linking to other datasets, such as the American Community Survey tables, utilizing the FIPS geolocation standard. In addition, the team resolved most null county values for the relevant installation records using available city information, as well as pulling up the utility service maps for utility-entered records with null county values, such as Tucson Electric Power, which primarily services Pima County, Arizona. As the team progressively resolved null values, it performed targeted updates of the Tracking the Sun table. Unresolved null values ultimately represented <1% of all relevant installations.

2. Product price: Given the irregularity and small sample sizes of many county installation records, particularly by quarter, the team decided to use the Lawrence Berkley National Laboratory methods described in its annual publication Tracking the Sun to estimate average cost per installed watt nationally rather than at the county-level for residential PV systems. This method suppresses the outsized impact of collection error in small sample sizes and what the team observed as outsized fluctuations in installed cost per watt for some counties quarter-to-quarter. While installation costs for PV rooftop systems are arguably at least partially driven by local factors, the team concluded that the perceived measurement error outweighed the benefit of greater granularity. Also, given that other county-level features might at least partially capture the higher cost of construction services and could partially compensate for this lack of low-level data, the team decided that the model could potentially still function reasonably well under these data conditions. National average price per installed watt was estimated quarterly and then joined with the instances using year and quarter as the join criteria.

3. County-level housing stock and income data: The American Community Survey (ACS) allows users to aggregate housing stock and income data at the county-level with FIPS geolocation data in its data request tool. Given that the team had already associated its instances to FIPS geolocations, it was a simple series of SQL joins to bring in the desired housing stock, household, and income data to the instances and features table. That data represented single-family homes, townhouses, households, and detailed income breakdowns by county estimated over the previous 5 years to 2015. Given the relative static nature of housing stock and incomes over the short- to mid-term, the team chose not to use multi-year data. The primary purpose of the housing stock and income features was to differentiate between counties and their relative demand for PV rooftop systems and not on the temporal component of that variation.

4. Average retail price of electricity by state (2015): The team then joined in the average retail price of electricity by state into its instances / features table. Based upon the relative stability of electricity prices over time, the team again chose not to incorporate a temporal aspect to this feature, but rather used only 2015 prices to differentiate between markets and the relative attractiveness of PV rooftop installations between geolocations.

5. State policy and regulation grades (2015): The team then linked in Freeing the Grid grades of state policy and regulations with respect to solar using state abbreviations as its join criteria. Those grades were then translated into numeric grades (A = 1, B = 2, etc.) using SQL query logic for model estimation purposes.

6. Presidential Election Results 2012: The democratic share of votes derived from the Presidential Election Results 2012 data by county were also linked to FIPS geolocation codes and then joined into the instances / features table using the same logic detailed above for housing stock and income data.

7. Latitude data: The 2015 National Counties Gazetteer File contained basic geographic data for each county, to include latitude data and county-level FIPS geolocation codes. Using the FIPS geolocation codes as the join criteria, the team extended the instances / features table to include latitude.

8. Solar Insolation data: Using the CDC WONDER Data request tool, it was possible to download average solar insolation data for the prior 7 years by county and associated FIPS geolocation code. The team was able to use the FIPS code to join the solar insolation data by county into the features / instances table.

Bringing the data together, as detailed above, involved successively joining data into a common instances / features table that the team was then able to download into a csv file and pipe into the various scikit-learn machine learning algorithms.

## IV. Machine Learning and Feature Transformations

Utilizing a series of Python pipeline scripts, the team proceeded to pipe in the csv file with all of the instances and features that we had collected into the machine learning Python scikit-learn models, including:

1. Ridge regression
2. Randomized Lasso
3. Perceptron
4. Support Vector regression
5. Linear regression
6. Random Forest
7. Decision Tree Regressor
8. K-neighbors regression

Initial results from the linear regressions were of low explanatory value. However, all of the non-linear regression models showed high explanatory power, with R-squared values hovering around 0.95. Although the non-linear results showed strong promise in creating a powerful predictive model, concerns were raised about overfitting. Also, the poor results from the linear regressions raised concerns: the economic theory of demand is well-supported with empirical evidence, often utilizing simple linear regression techniques to measure the validity and magnitude of the theoretical determinants of demands. The team thus decided to engage in feature transformations as a way of improving the linear regression model performance in a way that would aid the team in explaining the why and how these features are important to predicting PV rooftop demand.

Key feature transformations included:

1. Dividing the average retail price of electricity by the average installed price per watt as a measure of the relative price of the solar PV rooftop system to its most prevalent substitute: utility-produced grid electricity. This new variable was then defined as the grid-to-panel cost ratio. Furthermore, the exponent of this ratio was then multiplied by the sum of single-family homes and single-unit townhomes to account for access to the complementary goods within the county: rooftops. The team used the exponent based on the idea of a tipping point, whereas once PV rooftop systems approach and surpass grid parity in a given county / market (the point at which the economic break-even point is reached), it becomes non-linearly more attractive - from an economic perspective - to install a PV rooftop system with each successive drop in installed cost per watt.

2. Multiplying the democratic share of the 2012 presidential vote by the sum of single-family homes and townhomes to weight the variable appropriately to the availability of the key complementary good: rooftops.

3. Taking the exponent of the net metering and interconnectivity state policy grades and multiplying the results by the sum of the relevant single-family and townhouse counts.

After performing the above feature transformations, piped the newly transformed dataset into the scikit-learn algorithms and saw a leap in the explanatory power of the ridge and linear regression models from < 0.3 to 0.6 – 0.73, lending more support to the non-linear models' explanatory power. Given the high explanatory power previously observed in the non-linear models absent the feature transformation, it was unsurprising that the team did not observe a significant change in the explanatory power of the non-linear models using the transformed features. The team then plotted the expected and predicted values of the quarterly installation counts for each of the non-linear models and performed cross-validations. Based on visual observation, it appeared clear that the Random Forest regression model suffered significantly less from heteroscedasticity than the other models. While each of the models suffered from observable increases in variance as the predicted and actual values of the installation counts rose, the Random Forest model by visual inspection was clearly less impacted by this phenomenon than the other non-linear models.

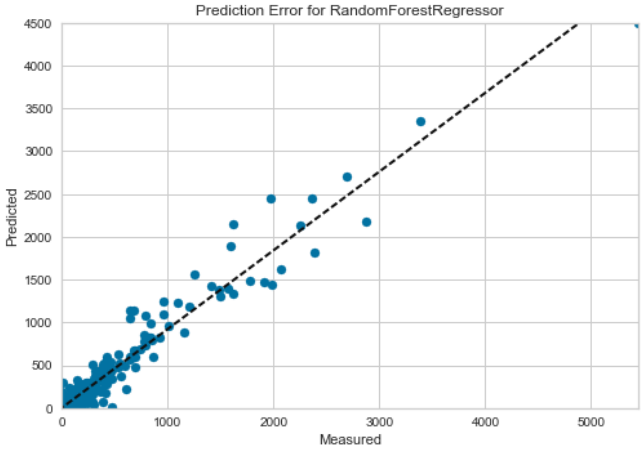


Fig. 1. RandomForestRegressor

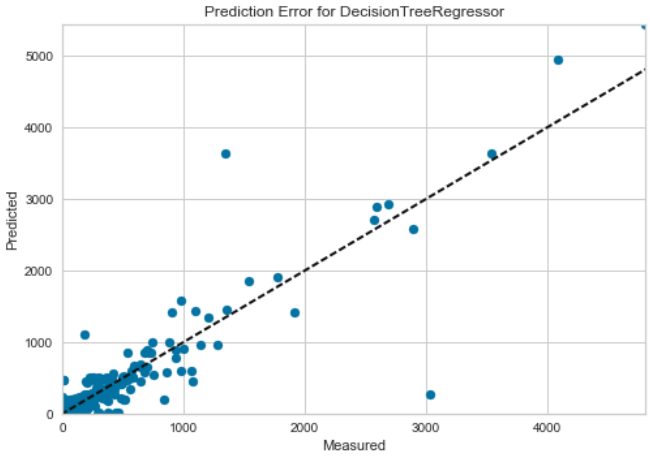


Fig. 2. DecisionTreeRegressor

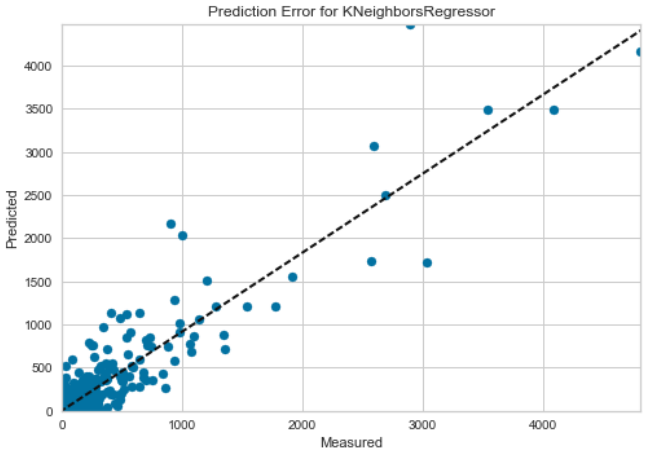


Fig. 3. KNeighborsRegressor

The cross-validation scores across the non-linear model set also reinforced the conclusion that the Random Forest model estimates were superior to the other models tested. With 12 folds of the data, the cross validation scores for the above models were:

- Decision Tree Regressor: 0.55
- Random Forest Regressor: 0.63
- K-neighbors Regressor: 0.46

Another reason for favoring the Random Forest regressor model over the others is its greater resistance to overfitting and bias. The ensemble approach of the Random Forest model reduces its relative susceptibility to this type of error over other non-linear models, which bears out in its superior cross-validation results.

Based on the model results, the team pickled the Random Forest regression model results for use in its data product that would predict PV rooftop installations based on key user inputs, such as price (development) and time.

## V. Data product using the random forest predictive model

To predict future quarterly installation counts per county, the pickled random forest regression model requests the values of the following 12 features as inputs:

- X1: sft_qty (number of single-family house)
- X2: th_qty (number of attached house)
- X3: households (number of total households)
- X4: hhincome_75_below_100k (number of households that has an income between 75k and 100k)
- X5: hhincome_100_below_150k (number of households that has an income between 100k and 150k)
- X6: hhincome_150_below_200k (number of households that has an income between 75k and 100k)
- X7: hhincome_above_200k (number of households that has an income above 200k)
- X8: dm_share* (sft+th) (Democratic share*(number of single_family house + attached house))
- X9: exp(interconnectivity) * (sft+th)
- X10: exp(net_metering) * (sfh + th)
- X11: exp(grid_to_panel) * (sfh + th)
- X12: latitude

During machine learning analysis, the team used the randomized lasso algorithm to evaluate the importance of all the 12 features. Result shows that the "Exp(grid_to_panel) * (sfh + th)" (feature X11) has the highest feature score (1.00 out of 1.00). This feature is calculated as the exponent of the ratio of the relevant electricity price to installation price multiplied by the sum of single-family and attached houses in that county. Theoretically, the electricity price and the house market would stay relatively stable over short- to mid-term while the installation price drops significantly. (For instance, the price dropped at an average rate around 9% per year over the last 8 years.) In addition, the values of the other 11 features are also considered relatively stable over time. Therefore, the team suggested that the installation price is the key factor that affects the installation counts, and therefore, it is the primary variable requested from the user for forecasting the installation counts per county.

More specifically, the final data product API works in the following steps:

1. The installation price is requested as an input and the value for feature X11 "exp(grid_to_panel) * (sfh + th)" is then calculated automatically
2. Feature X11 along with the lasted values (mostly in 2015) of the other 11 features gets plugged in the pickled random forest model for predictions.
3. A choropleth map at the county level is then generated as the data product's visual output.

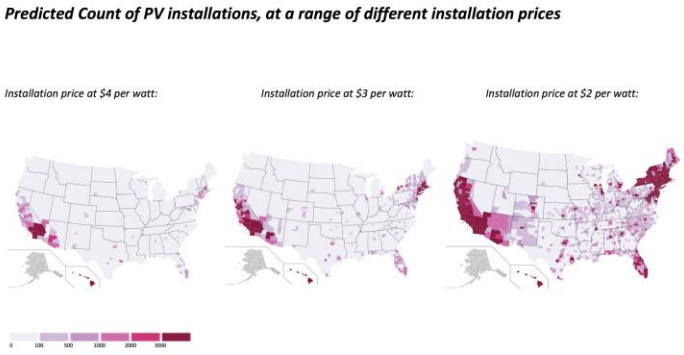


Fig. 4. The visualization maps above indicate the concentration of growth that would be anticipated as the installation price decreases, particularly in Florida, the Southwest, Mid-Atlantic, and the Northeast. This prediction results also illustrate the strong impact of installation price changes on the predicted quarterly installation by county.

However, for the more unsophisticated user who does not have a reasonable basis for estimating the future price per watt, the team designed two other options for predicting quarterly installation counts:

1. The user can input an expected rate of price decline, and a year or range of years. By using the latest installation price data (in 2015) as a baseline, the estimated installation price in a specific year can be computed and then passed to the API for prediction. This approach requires the user to have a greater insight into potential technological developments and related impacts to price.

   As an example, if we assume the price drops at a rate of 9% per year (which is the average rate of price decline over the last 8 years), the predicted installation counts per county from 2016 to 2021 would render the following growth pattern.

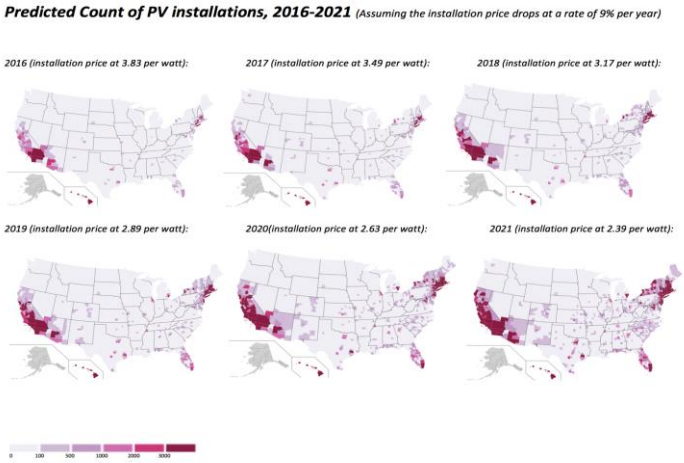


Fig. 5. The visualization maps above indicate the concentration of growth that would be anticipated as the installation price decreases at an assumed rate of 9% per year

2. The team also trained a simple price model using a time variable (year and/or quarter) as the only feature, exploring multiple model sets such as linear repressor, logarithmic and non-linear regression (decision tree, random forest). After comparing each model's R squared score and predicted results for the next couple of years, the team picked a simple logarithmic model, which has the highest R squared score (0.86). The predictive equation from the model is: installation_price_per_watt(Y)= -6.358*ln(X) +25.602. Using this model to translate a user's input of year or quarter into a predicted price, the data tool then passes the predicted price to the API for the prediction of quarterly installations counts by county

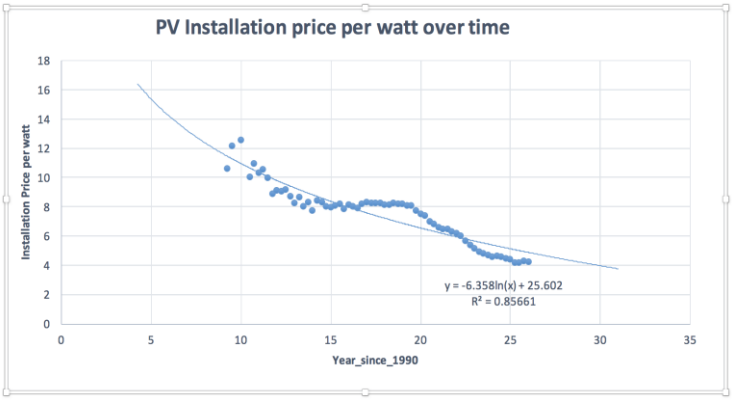


Fig. 6. A logarithmic model of installation price per watt rates over time.

This model provides the data product user a simplified alternative for estimating the growth of PV installation over the long term. However, the selected model tends to overestimate the installation price in recent years. Although the equation produces a reasonable estimation for the national average price, the prediction error is magnified and compounded due to the exponential function embedded in feature X11 (Exp(grid_to_panel_ratio) * (sfh + th)"). For a more precise prediction of installation counts, the team recommends that users input their own estimates of price per watt for short-term (1-5 year) predictions. Additionally, the user can select a geolocation (county, etc.) for more specific prediction results.

## VI. API DEVELOPMENT AND OPERATIONALIZATION

The resulting model of the aforementioned analysis, feature engineering and subsequent model development is then developed and exposed as a web application. Remaining consistent with the technologies used for data ingestion and exploratory analysis, the application architecture leverages a python based open-source Model-View-Controller (MVC) architecture called Django [9]. The data architecture is as follows:

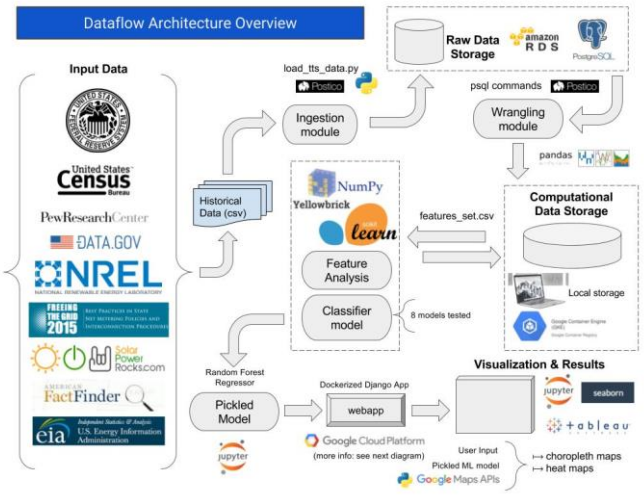


Fig. 7. A diagram depicting the logical flow of data and computation for the data product.

Targeted users would be domain experts who are interested in discovering new markets, or who may be curious about market potential given product cost differences. Said user could refer to Fig 6. to guide their forecasting parameters (installation price/watt in respect to time) and see choropleth and heatmap visualizations depicting counties predicted to experience high demand.

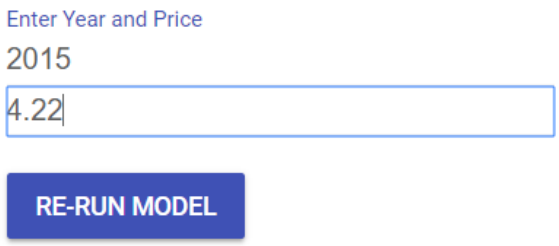


Fig. 8. A screenshot of the input interface on the website that exposes the API – prompting the user to enter a year and installation rate ($/watt) that will be used as parameters in the predictive model.

The resulting visualizations are geocoded latitude and longitude pairs from the resulting counties in the predicted set, and used as plotting parameters for the Google Maps JavaScript API [10].

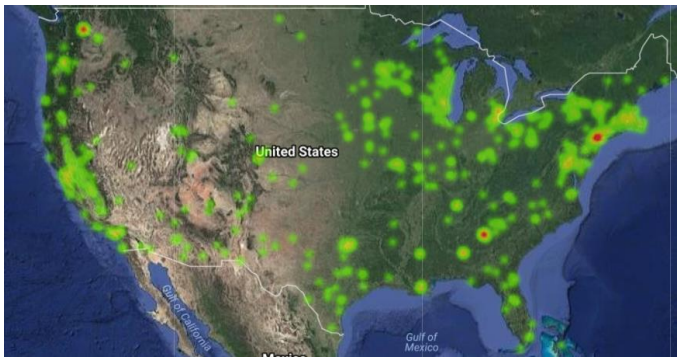


Fig. 9. An unweighted heatmap visualization of the 500 counties in the contiguous United States predicted to have the most PV installations at a $2/watt installation rate

The deployment of this model gives researchers, market competitors and enthusiasts access to a carefully crafted predictive tool.

## VII. Recommendation for Future Work

The team's first recommendation towards improving the data product would be to improve the underlying predictive model. As discussed previously, the team explored a number of machine learning algorithms in building the predictive model it selected, and the team picked the best performing model – a random forest regression model. Although the model achieved a high R square score (0.95) and achieved good, if not great, cross-validation performance scores, it does not perform as well in estimating values at the extremes (with a problem of underestimating high values). To accomplish this, the team would ingest and incorporate other datasets to study their impacts on the predictive model. Potential datasets worth exploring would include county level tree canopy data, previous install bases of PV system, and Yelp review data for local PV rooftop installers. Another avenue worthy of pursuit would be to transform the label data (independent variables) to make it more centered. This effort might improve the model's performance on predicting values at the extremes.

The data product's secondary model for predicting installation price over time was trained using national level data. To improve this model, we would potentially ingest and incorporate county level data to enable a more localized prediction of installed price per watt. The team would also consider incorporating other datasets that affect the installation price temporally, such as variables that predict the advancement of PV technology or track utility / local government incentives / rebates. other datasets that affect the installation price temporally, such as variables that predict the advancement of PV technology or track utility / local government incentives / rebates.

## VIII. Conclusion

Photovoltaic solar cell technology is a rapidly advancing source of renewable, inexhaustible, and clean energy for utilities, businesses, and households. Among alternative energy technologies, the rooftop photovoltaic (PV) system is one of most popular and increasingly attractive technologies available for residential use. Over the past decade, driven by the rapid development of revolutionary PV technologies and drops in overall installation price, rooftop PV installations have experienced exponential growth in the US and around the world. The market is projected to continue its rapid growth, driven by continued improvements in PV cell efficiency, installation cost improvements, and society's concerns about the detrimental impact of carbon release associated with traditional energy generation.

## IX. References

*[1]* PV rooftop installation data collected by the National Renewable Energy Laboratory (NREL) via public contributions and through the Lawrence Berkley National

Laboratory (LBNL), available at https://openpv.nrel.gov/search

*[2]* PV rooftop installation data collected and cleaned by the Lawrence Berkley National Laboratory (LBNL) in their Tracking the Sun Public Data File available at https://openpv.nrel.gov/search

*[3]* County-level income and housing stock data collected in the American Community Survey by the US Census Bureau and available at https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml

*[4]* Average state electric utility prices for 2015, collected by the US Energy Information Agency and available at https://www.eia.gov/electricity/state/

*[5]* County-level Presidential General Election Results – 2012 available at https://www.theguardian.com/news/datablog/2012/nov/07/us-2012-election-county-results-download#data

*[6]* A-F grades of states' net metering policies and interconnection regulations / procedures produced by the Interstate Renewable Energy Council (IREC). http://freeingthegrid.org/#state-grades/

*[7]* 2015 National Counties Gazetteer File with latitude data by county available at https://www.census.gov/geo/maps-data/data/gazetteer2015.html

*[8]* County-level solar insolation data collected in the North America Land Data Assimilation System (NLDAS) Daily Sunlight (KJ/m²) (1979-2011) by NASA and available at https://wonder.cdc.gov/NASA-INSOLAR.html

*[9]* Django: The web framework for perfectionists with deadlines, https://www.djangoproject.com/

*[10]* Google Maps JavaScript API, https://developers.google.com/maps/documentation/javascript/