



# Analysis of Solar Power Cost over Time

## Final Report, Capstone Project 1

Springboard, DSCT

25Apr2018

---

Mark Sausville

[saus@ieee.org](mailto:saus@ieee.org)

## The Problem

The goal of this project is to produce a model that will allow prediction of the Fair Market Value (FMV) cost of installing a solar electrical system on a typical home in the 4-8 quarters beyond the end of the available historical data, so the customer can make an informed decision about whether to install now or defer to enjoy more favorable pricing.

## The Customer

The customer is a householder considering the installation of a solar electrical system. The question to be answered is “Will it be more cost-effective to install now, or will I save money by waiting a year or two?”. This is a reasonable question given the market context of rapidly falling costs for solar power installations over the past decade.

## Project Structure

This project breaks down into the following steps:

- Acquire the data.
- Identify and investigate the relevant data in the dataset and wrangle into a format that can be used for the following steps.
- Perform a statistical learning analysis to find a function that expresses solar installation cost as a function of time (and other relevant features) and use that function to predict the cost of installation in the near term. Compare the performance of these various models and determine how they will be used to generate predictions. Also, study the effect of geographic location in the modeling.
- Use the model to predict the FMV cost/watt of solar installation quarterly for two years beyond the end of the data.
- Make recommendations based on insights gleaned from exploration and modeling of the data.

## Data Acquisition

The data for this project were acquired at the site: <https://openpv.nrel.gov/search>. This site contains data sets from both the National Renewable Energy Laboratory (NREL) and Lawrence Berkeley National Laboratory.

Both datasets are of similar size and nature (around 1 million rows representing distinct installations, approximately 80 features). The documentation on the site implies that the NREL data is more comprehensive, so we began with that dataset.

After an initial cleaning of the NREL dataset (detailed in this [report](#)), we discovered that there was no mechanism in this data to prevent duplicate entries. In fact, many suspected duplicate entries were found. As there was no differentiating field available for these records, this dataset was set aside as unreliable.

Fortunately, the LBNL data was available and better structured, including unique identifiers and the source of the data for each record. The LBNL data also had coverage for 2016, while the NREL data ended with 2015. The LBNL dataset included documentation including a short description of each feature. The details of the data wrangling for this dataset are available [here](#).

## Cost Model for Solar Installation

In market cycles we have come to expect that as the number of adopters grows, the cost of adoption drops as economies of scale come into play (producer cost/unit falls) and competition forces producers and installers to operate more efficiently.

In technology cycles, these effects are often exaggerated by innovation and discovery. Moore's Law is a well-known example: for many years processor power/cost has roughly doubled every 18 months.

In solar energy, there is an analogous formulation, Swanson's Law, i.e., the price of solar photovoltaic modules tends to drop 20 percent for every doubling of cumulative shipped volume. While not as dramatic a decrease as Moore's Law, drops in the cost of solar panels have been sufficient to have cut the price by 50% in the last decade.

For the customer, if the installation price is falling rapidly, waiting for the price to drop may be beneficial. On the other hand, if price is flat or increasing, buying now may be the best option. This study aims to educate that decision by predicting near term price by extrapolating from a model learned from the data.

The cost of a solar installation has many components. The costliest single component has historically been the photovoltaic modules, but non-module costs currently amount to 1-2 times the module cost.

The following components of the price could be considered in an analysis:

- Solar Panels
- Other hardware
- Inverter
- Racking
- Cables, conduit, interconnection hardware, etc.
- Installation labor
- Installer/Integrator costs and profit

Unfortunately, the data did not support a detailed hardware/overhead breakdown and separate modeling of each component. While features existed in the data that seemed to allow this kind of approach, on examination these fields were largely unpopulated (or completely null).

LBNL explained that while the fields were documented, they had been unable to acquire reliable data for these features and had dropped them from the public dataset.

While this was disappointing, there is a strong basis for modeling the cost of solar installations. This will be apparent from an exploration of the data.

## Exploration of the data

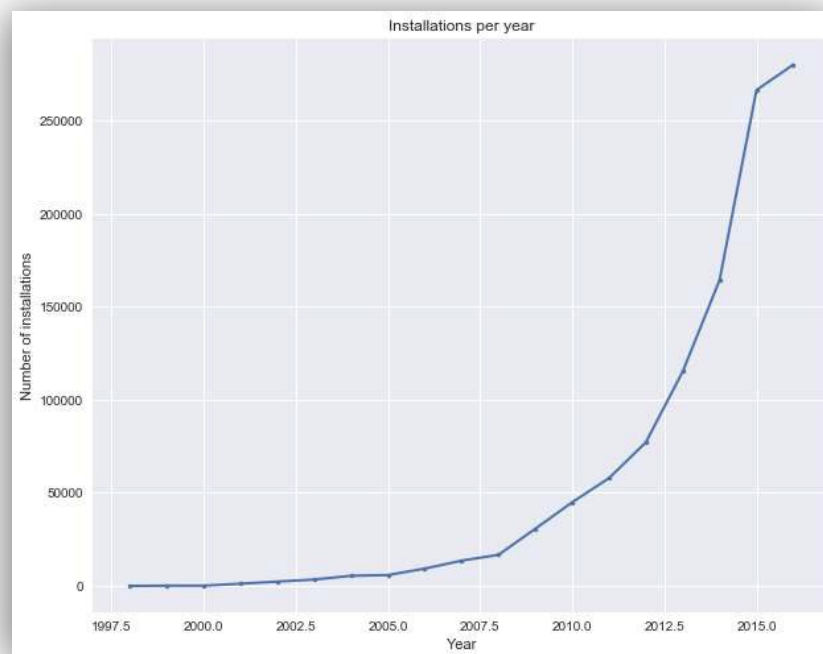
### The growth of solar power

We often hear that the number of solar installations has grown dramatically. The first question is: “Does the dataset support that claim?”.

While this dataset by no means describes every solar installation in the US (only 19 states are present in the data), all the largest state markets are included. The data are sourced primarily from state agencies and utilities that administer solar incentive programs, solar renewable energy credit registration systems, or interconnection processes.

The growth shown in Figure 1. is exponential through 2015, with an average annual growth rate about 50%. In 2016, the data show that growth rate may be beginning to decline. In any case, it is evident that US installations can be expected to continue in the hundreds of thousands for the near future as solar energy becomes more price-competitive with traditional power sources.

**FIGURE 1. NUMBER OF SOLAR INSTALLATIONS PER YEAR**



### What is the behavior of cost/watt over time?

The cost of a solar power installation is generally measured in dollars per watt and calculated by dividing the total size of the installation in watts by the total cost in dollars. As the media tell us, the price of solar power has decreased dramatically over the past few years.

FIGURE 2. MEDIAN COST/WATT IN RESIDENTIAL AND COMMERCIAL INSTALLATIONS

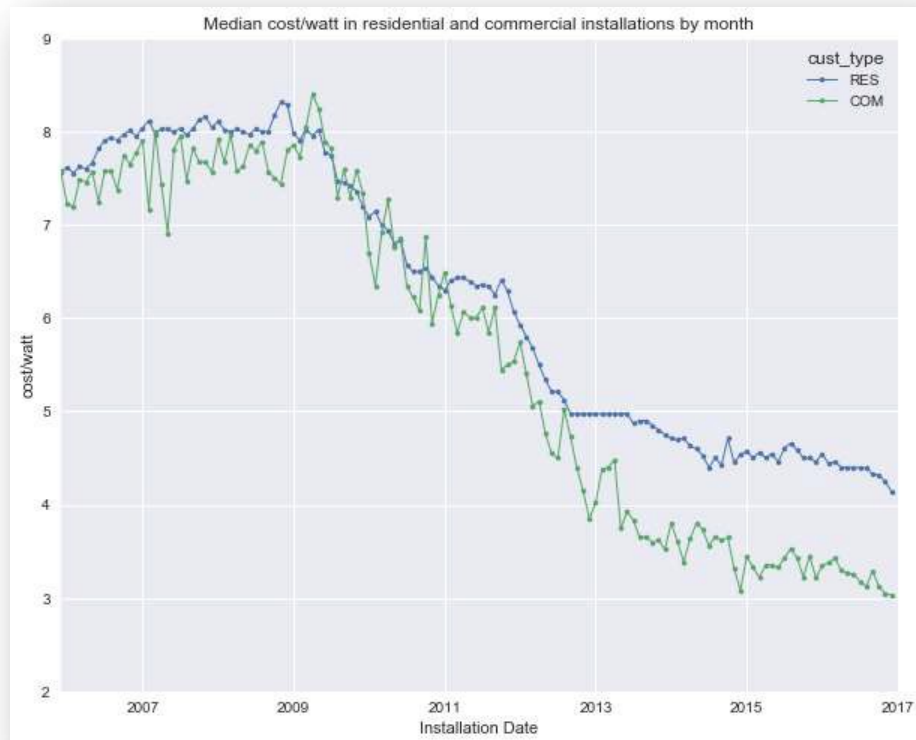


Figure 2 shows that cost/watt has fallen dramatically over the past decade. From the beginning of 2007 to the end of 2016, the median cost/watt for residential customers fell from about \$8/watt to near \$4/watt. Commercial customers have consistently paid less over this period showing a more pronounced reduction in price.

### Does cost/watt vary by customer type

We see in the graph above that commercial customers pay less per watt of solar capacity than residential users. Does that apply to the other customer types? There are seven types of customer in the dataset. Residential installations comprise 96% of the data. Commercial installations are 2%. The other five types represent only 2% of the data. Table 1 shows the number of each customer type and percentage of the total.

TABLE 1. NUMBER OF CUSTOMERS OF EACH TYPE

<i>Customer Type</i>	<i>RES</i>	<i>COM</i>	<i>NON-RES</i>	<i>GOV</i>	<i>NON- PROFIT</i>	<i>SCHOOL</i>	<i>TAX- EXEMPT</i>
<b>Count</b>	<b>745688</b>	<b>15199</b>	<b>7042</b>	<b>3996</b>	<b>1951</b>	<b>1748</b>	<b>70</b>
<b>Percentage</b>	<b>96%</b>	<b>2%</b>	<b>1%</b>	<b>1%</b>	<b>0%</b>	<b>0%</b>	<b>0%</b>

In Figure 3, we display the median cost/watt by month for four installation types. One thing to keep in mind while considering the graph: the number of nonresidential installations is quite small compared to the residential installations (30k vs. 750k).

Further, those 30k nonresidential installations are split among 6 types, the largest of which is commercial (business) installations. The smaller customer types show much more variability than the larger groups. This explains the peaks seen below. Only the residential and commercial customer types have enough participants for their aggregates to be well-behaved.

FIGURE 3. MEDIAN MONTHLY COST BY TYPE



Figure 3 shows that residential customers generally pay more at any given time than non-residential customers (see also Figure 2). One possible explanation for this difference is that nonresidential customers typically buy larger installations and therefore get more favorable pricing. We examine the relationship between cost/watt and size of installation below.

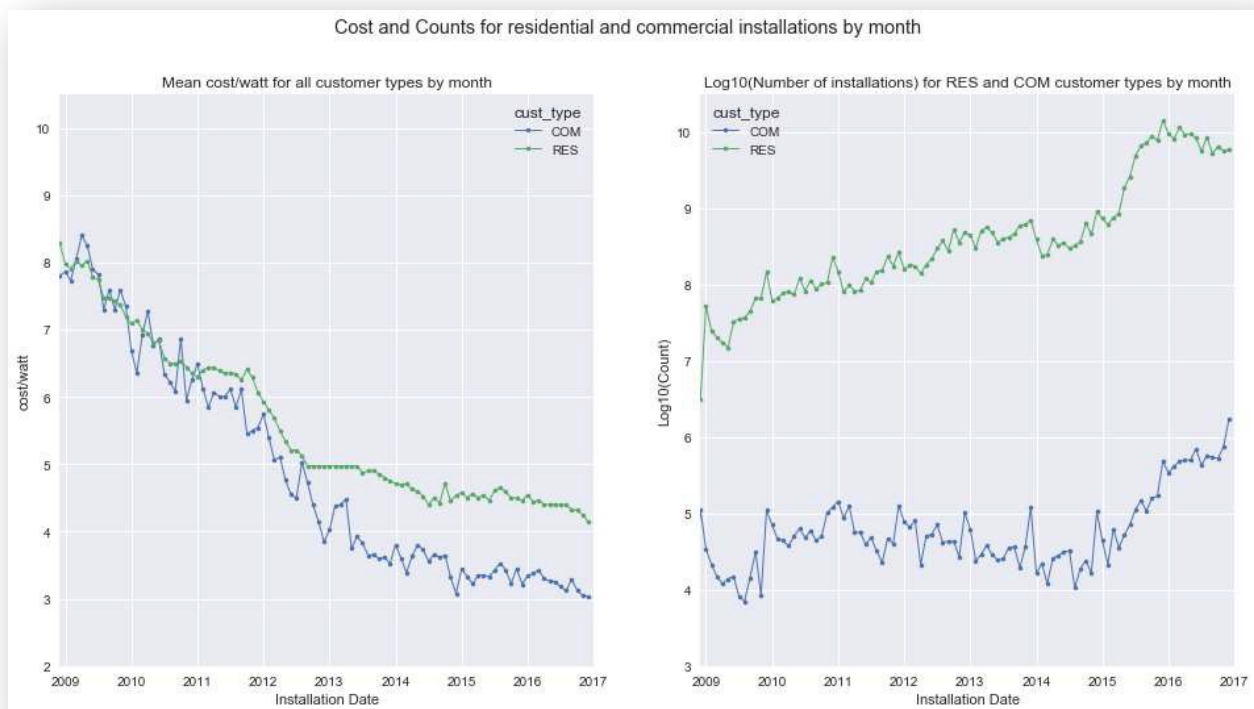
### Does cost/watt decrease with the size of the installation?

We see above that residential installations tend to have a higher cost/watt than all other types. Table 2. shows that residential installation median size is 5.75kw, while commercial installations are generally larger with a median of 27.72kw.

TABLE 2. MEAN AND MEDIAN SIZE AND COST/WATT FOR RESIDENTIAL AND COMMERCIAL SOLAR INSTALLATIONS

	<i>Com size (kw)</i>	<i>Com cost/watt</i>	<i>Res size (kw)</i>	<i>Res cost/watt</i>
<i>median</i>	25.2	4.7	5.75	4.94
<i>mean</i>	131.3	5.21	6.45	5.14
<i>count</i>	15199	15199	745688.	745688.

In Table 2, the mean (and median) cost does not appear to differ drastically over the full dataset between commercial and residential installation. That is because the bulk of residential installations come in recent years at lower prices, pulling down the mean and median cost. While the number commercial installations grew quickly starting in 2015, the residential installations were 1000 times greater. This is clearly visible in Figure 4 where the installation counts are shown on a logarithmic scale.



**FIGURE 4. MONTHLY MEDIAN COST/WATT AND COUNTS FOR RESIDENTIAL AND COMMERCIAL INSTALLATIONS**

It is reasonable to expect that commercial installations have less expensive per watt costs due to quantity discounts. We examine this assertion below.

To see if cost/watt varies with size, it is important to compare installations where only the size differs, so we restrict the set to residential installations. Residential installations comprise 96% of the data. We begin by looking at the distribution of sizes.



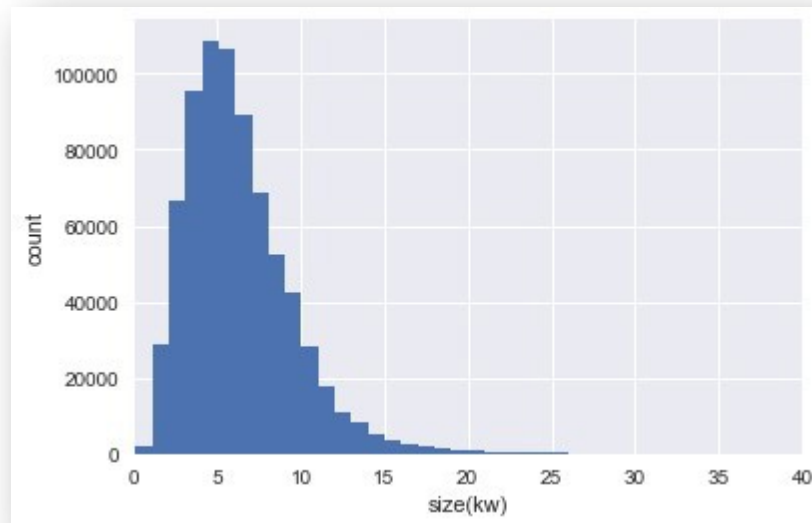


FIGURE 5. DISTRIBUTION OF SIZE (kW) FOR RESIDENTIAL INSTALLATIONS

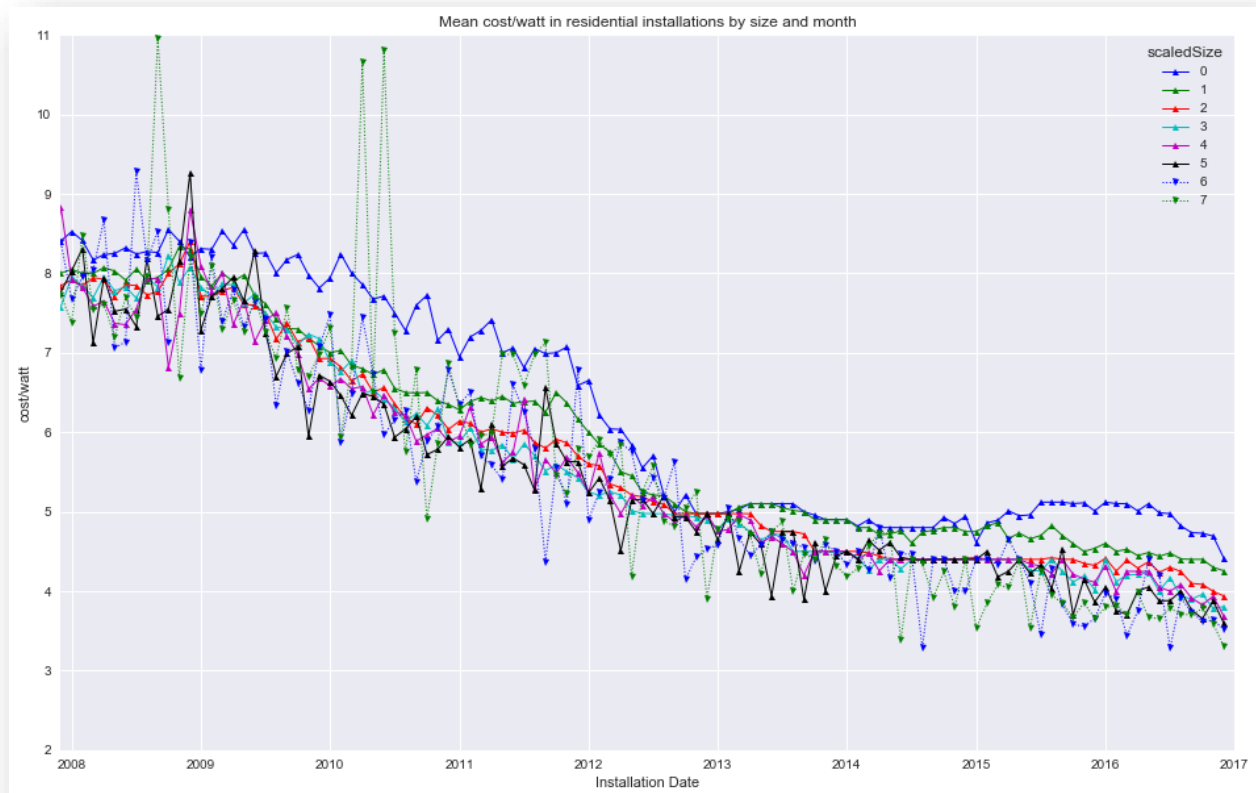
TABLE 3 SUMMARY STATISTICS FOR SIZE (kW) IN COMMERCIAL AND RESIDENTIAL INSTALLATIONS

	<i>count</i>	<i>mean</i>	<i>std</i>	<i>min</i>	<i>25%</i>	<i>50%</i>	<i>75%</i>	<i>max</i>
<b><i>Com size(kw)</i></b>	<b>15199</b>	<b>131.3</b>	<b>373.8</b>	<b>0.33</b>	<b>10</b>	<b>25.2</b>	<b>80.6</b>	<b>5999</b>
<b><i>res size(kw)</i></b>	<b>745688</b>	<b>6.45</b>	<b>7.1</b>	<b>0.11</b>	<b>4.03</b>	<b>5.75</b>	<b>7.95</b>	<b>1989</b>

Figure 5 shows the distribution of size in kilowatts for 746k residential installations. As the summary statistics show, 50% of the residential installations are between 4 and 8 kilowatts.

In the following figures, we plot cost/watt vs. time, distinguishing different size groups.

FIGURE 6. COST/WATT VS. INSTALLATION MONTH BY SIZE



The three graphs, Figs. 6, 7 and 8 all show the same data, depicted in three different ways. In each, the size of the installation is binned into 2.5kw-wide bins (i.e. 0-2.5 kw, 2.5-5 kw, etc.). Anything larger than 25kw is grouped in the bin representing the largest installations.

The first graph (Figure 6) is perhaps the clearest in that the groups are visually separable, and it is quite apparent that within most month-long periods, the smallest installations pay the highest price, with price decreasing as size increases.

The second two graphs add visual intuition about the shape of the surface of cost/watt as a function of both time and size, showing a surface with generally negative slope as both time and size increase. The 'cliffs' in the surface plot are due to filling in missing data in the size-group space. In the earlier data, few size groups are populated as the number of installations is smaller. At later dates, as the number of installations increases, all size groups are represented, and the surface becomes better behaved.

From smallest to largest, all groups show a decrease in cost over time. The decrease in price as size increases while time is held constant is not as steep. Note that in the surface plot (Figure 8), the darkest blue area (corresponding to least expensive cost/watt) is found at the latest time and the largest size group.

FIGURE 7. COST/WATT VS. TIME AND SIZE -PSEUDO-SURFACE PLOT

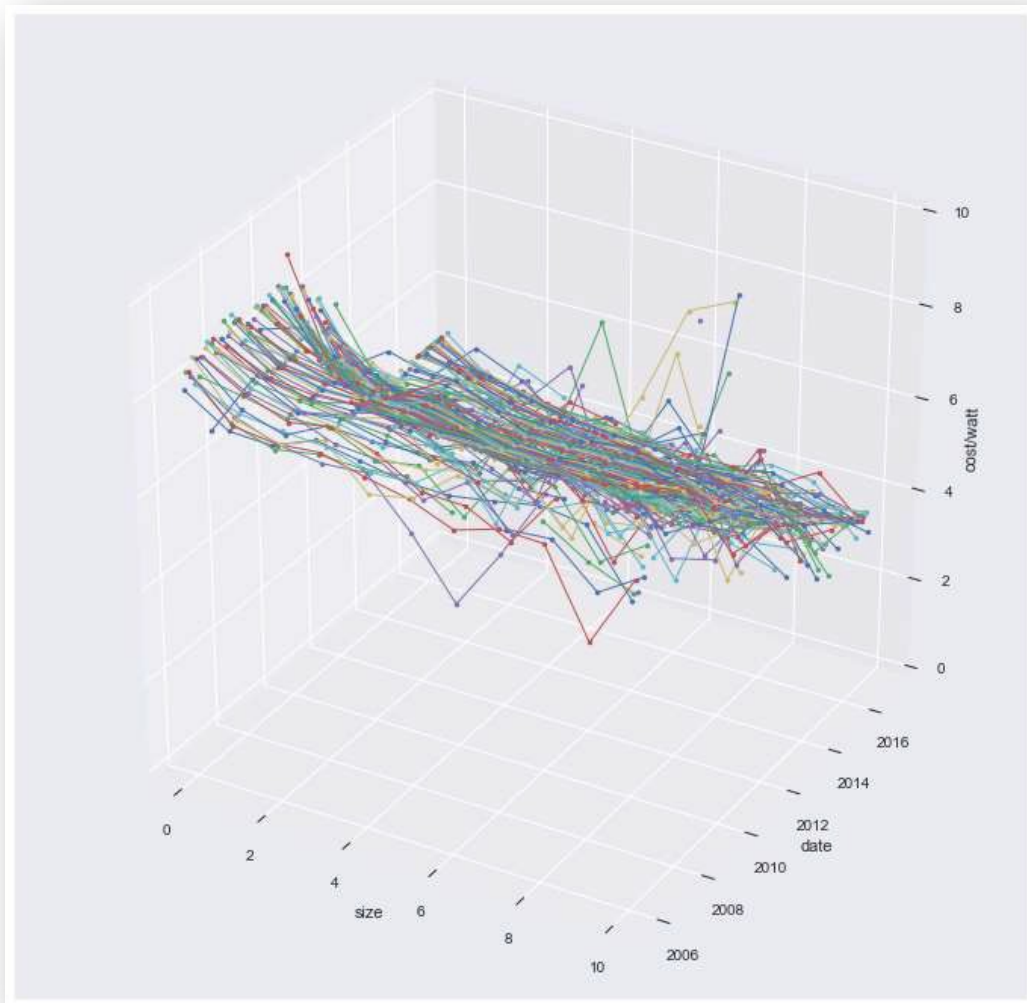
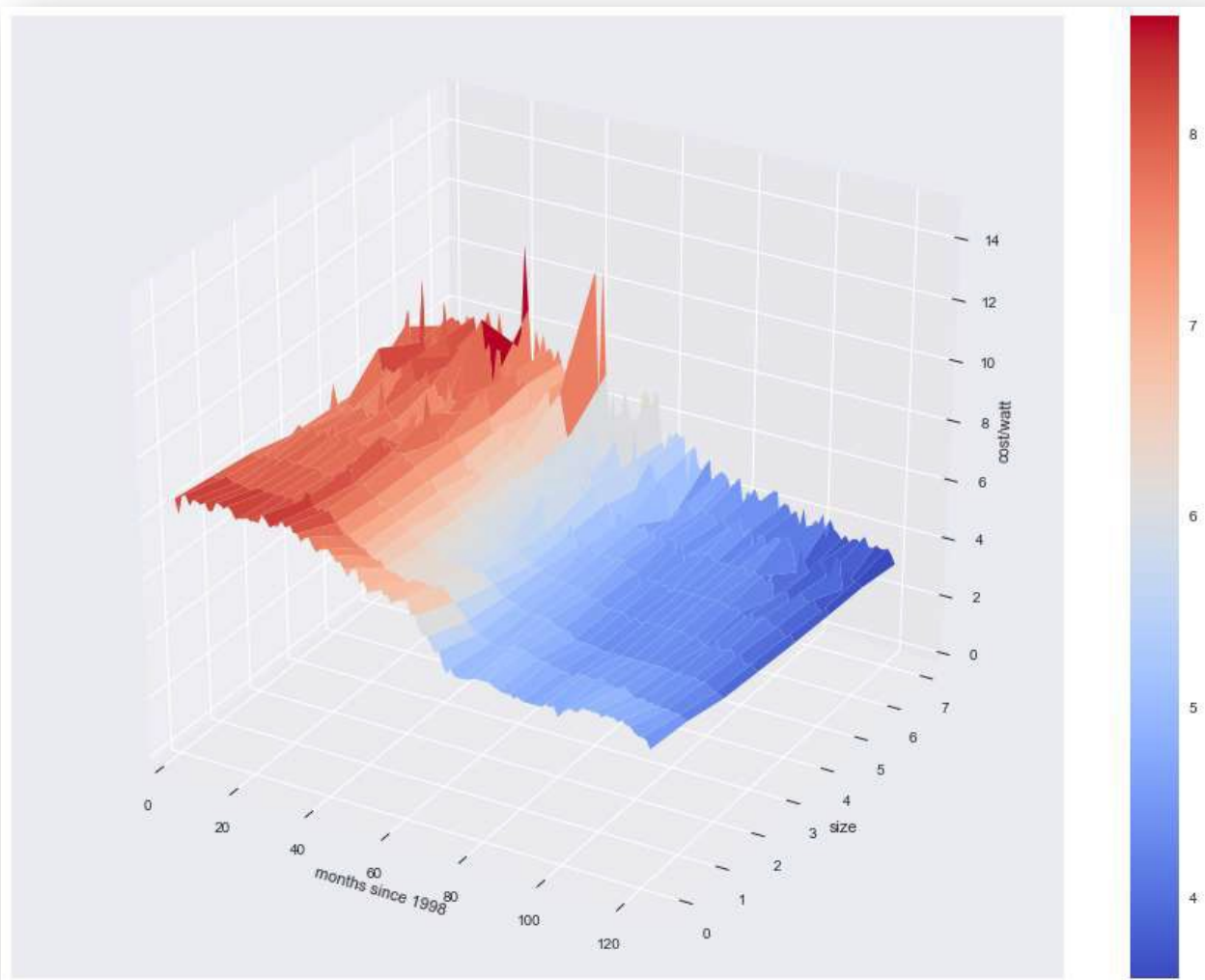


FIGURE 8. COST/WATT VS TIME AND SIZE – SURFACE PLOT



## Does cost/watt vary by region?

It is natural to ask if the cost of solar power varies by location. Figure 9 indicates that location has a significant impact on the market price of solar installations.

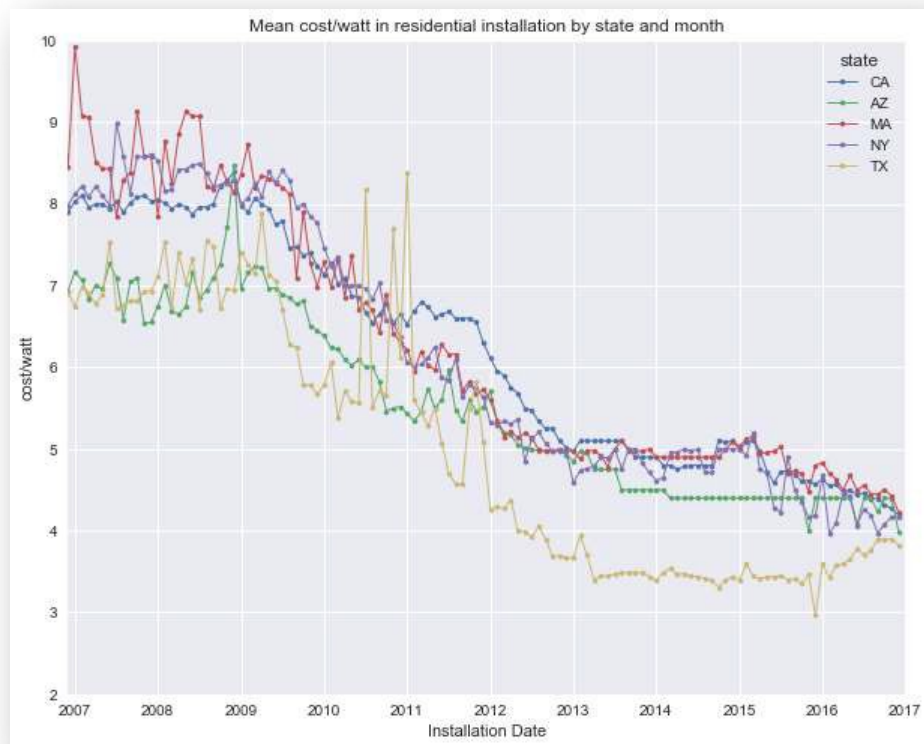
The differences in market price for solar installations can perhaps be explained by 'cost of doing business'. Texas is consistently the least expensive state. California, where higher costs are offset by strong incentives is by far the largest market. For visual clarity we show only the five largest state markets. California, Arizona, Massachusetts, New York and Texas together account for 90% of the US market.

Figure 9 has two interesting features:

Historically the cost of solar power has varied widely from state to state. For example, in 2014, the median cost/watt in New York was over \$5/watt, while in Texas the median cost was about \$3.50. In earlier years, prices diverged by even greater amounts.

Over time, the differential by state seems to be diminishing at the same time as costs decrease in almost every state.

**FIGURE 9. MEDIAN COST/WATT VS. TIME BY STATE**



## What is the impact of the Boolean features on cost/watt?

The following features are Boolean variables in the dataset. There are two different kinds of features: those having to do with price characteristics and those associated with hardware in the installation.

These features describe how the price in the dataset can be interpreted.

- `third_party` - Is the installation is owned by a third party?
- `appraised_value` - Is the cost/watt a market price or an appraised price?
- `new_const` - Is the installation part of the initial construction of the residence?

Third-party installations are owned by the vendor (i.e. neither the consumer or utility) and typically installed at no upfront cost to the consumer. The installation is funded by the owner of the system in conjunction with a long-term power purchase contract with the consumer.

An appraised price is not as trustworthy as a market price because it is possible to manipulate. There are financial motives to report an inflated or reduced cost appealing to various parties (installers, consumer, system owner, incentive program participants).

New construction may provide an opportunity to save on installation costs.

The features below also take the form of Boolean variables but represent physical aspects of the installation.

- `ground mounted`
- `battery`
- `tracking`
- `microinverter`
- `DC optimizer`

Ground mounted systems may be better located with respect to the amount of sun and may offer a simpler (and less expensive) installation. Battery equipped systems store power for use at night but are costlier. Tracking systems incorporate machinery to aim the solar array for higher power production but are costlier.

Microinverter systems are simpler to install and configure but cost more in aggregate than a single large inverter. DC optimizers provide advantages in flexibility of system design and management at additional hardware cost.

## Appraised Prices

It is clear from the data (see Figure 10) that appraised prices tend to exceed market prices for solar installations. One implication for an analysis is that since there are so many (approximately 230k) it may be worthwhile to exclude appraised prices from a modeling dataset.

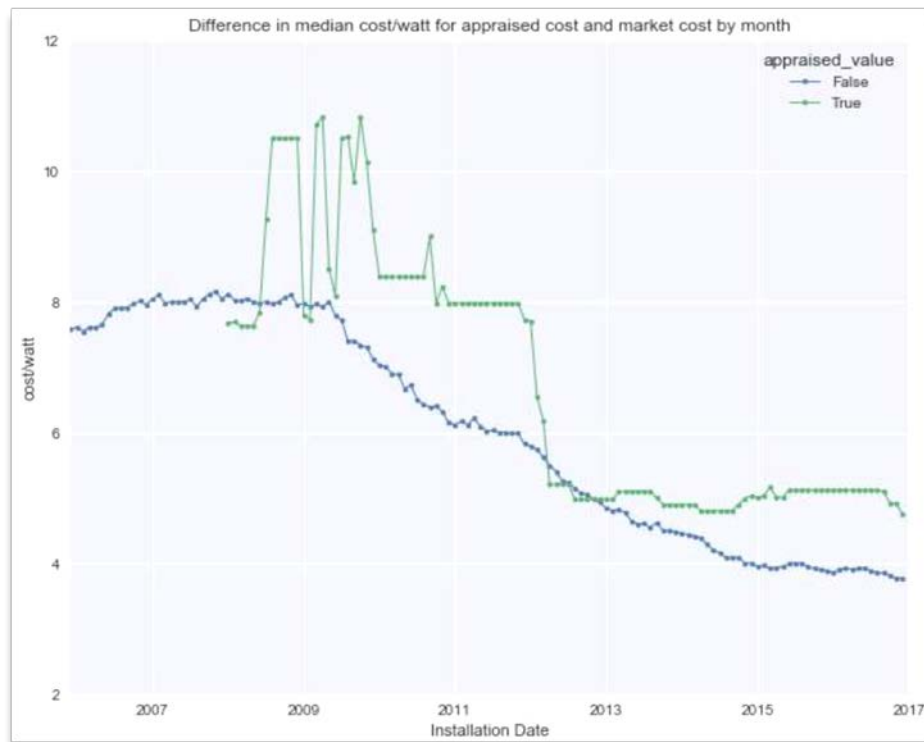


FIGURE 10. COMPARE APPRAISED AND MARKET PRICES

### Third-Party Ownership

It is interesting that in the 2006-2007 time frame (see Figure 11), third-party ownership becomes popular as a result of tax incentives and the wide availability of credit to third-party owners.

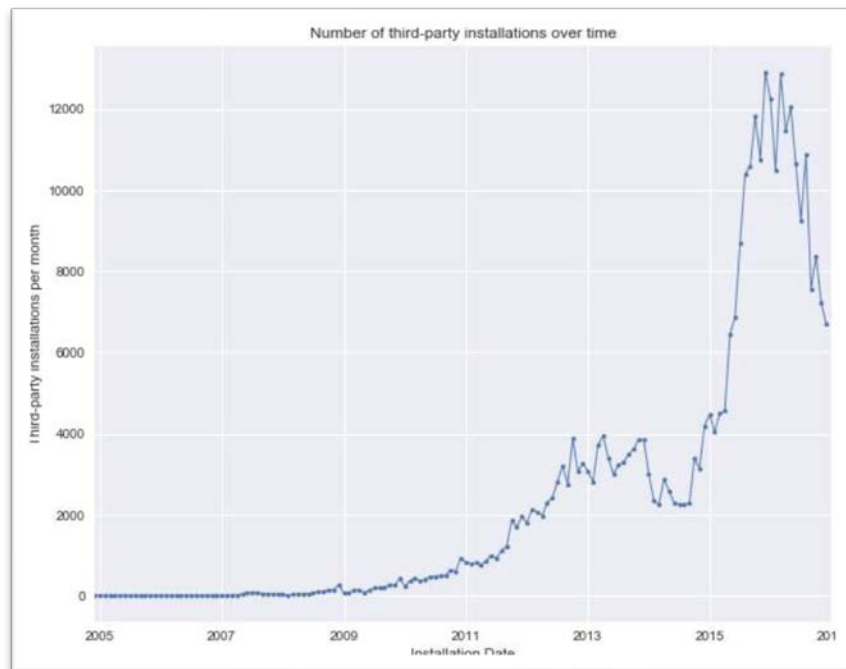


FIGURE 11. NUMBER OF THIRD-PARTY OWNED SYSTEMS INSTALLED VS. TIME

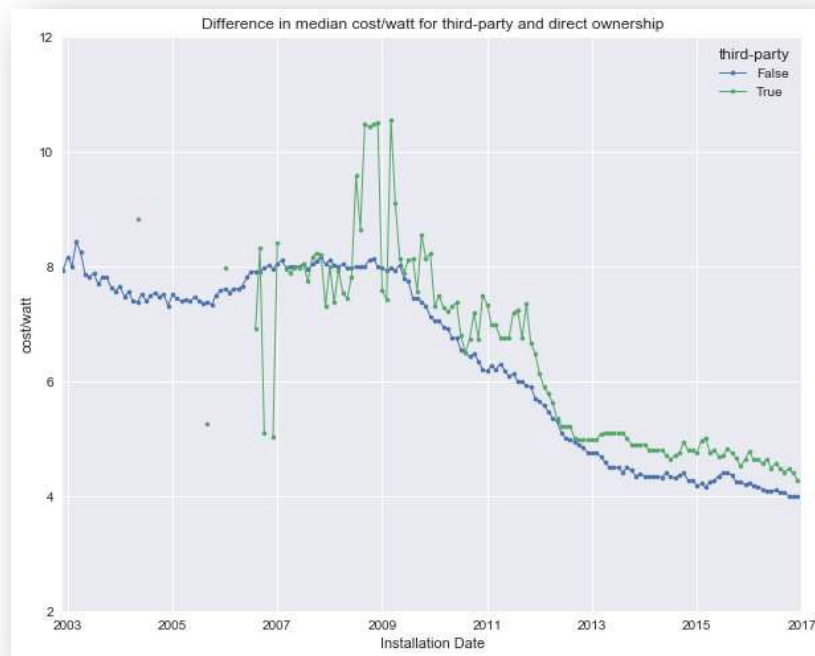
There seems to be a consistent trend for third-party pricing to exceed direct ownership pricing though the difference tends to be less in recent years (Figure 12).

There are many reasons for the differential (e.g. price may be inflated since third-party owner often controls the supply chain and installation process). We will not examine this in detail in this project.

It should be noted that third-party owned systems comprised a very substantial subset. For accurate analysis of market price, it may be necessary to exclude these from cost/watt modeling.



FIGURE 12. COMPARE THIRD-PARTY/MARKET PRICES



### Other Boolean features

The other Boolean features in the data were examined and found to be not relevant to the thrust of the project because a) they are not well represented (e.g. battery systems) or b) have little impact on price (e.g. DC Optimizer). For those interested, the examination of these features (as well as the rest of the EDA code) is presented [here](#).

## Statistical Modeling

### Introduction

The goal of the modeling effort was to create a model of cost/watt in solar installations that accurately captures the price trend in order to make quarterly price predictions for the two years past the end of the data (2017, 2018).

The EDA phase suggested that time, size and location all had an impact on the cost/watt of a solar installation. We used these features to model the cost.

### Features of the dataset

The dataset has some characteristics that impact modeling. For example, the bulk of the installations are in the last three years (see Figure 1). Specifically, the raw dataset has about 711,000 installations in 2014, 2015 and 2016. Contrast this with 384,093 total installations for the previous 16 years, 1998 through 2013. This means that grouping operations will have much larger groups in later years. Group aggregates for earlier years may have erratic results.

The dataset is also quite noisy. For example, Figure 13 is a histogram of prices paid in 2016 for installations in California between 4.5 and 5.5 kilowatts.

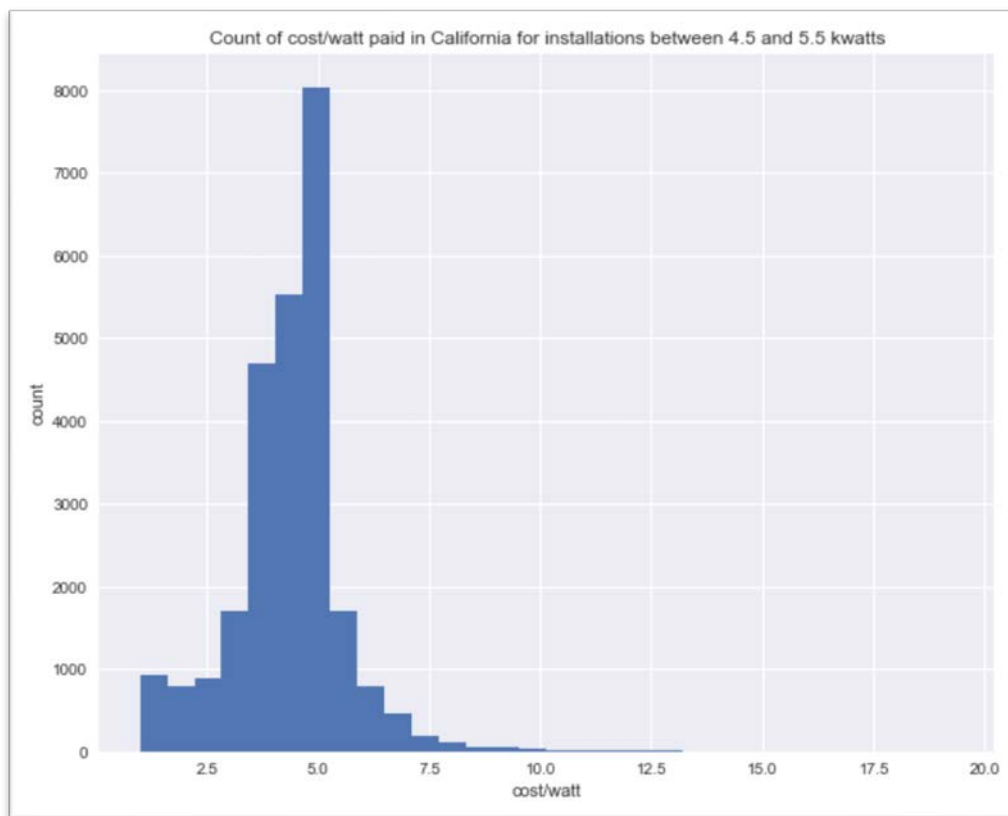


FIGURE 13. HISTOGRAM OF PRICES FOR SIMILAR INSTALLATION IN CALIFORNIA

TABLE 4 SUMMARY STATISTICS FOR COST OF SIMILAR INSTALLATIONS IN CALIFORNIA

	count	mean	std	min	25%	50%	75%	max
cost_per_watt	26107	4.41	1.33	1.00	3.74	4.50	5.12	19.30

Table 4 has summary statistics for this grouping of similar installations. There does not seem to be a basis in the data for the large variance in price. The lowest price among this group is \$1.00/watt; the highest is \$19.30. 50% of installations were priced between 3.75 and 5.12. That implies 50% are out of this central band.

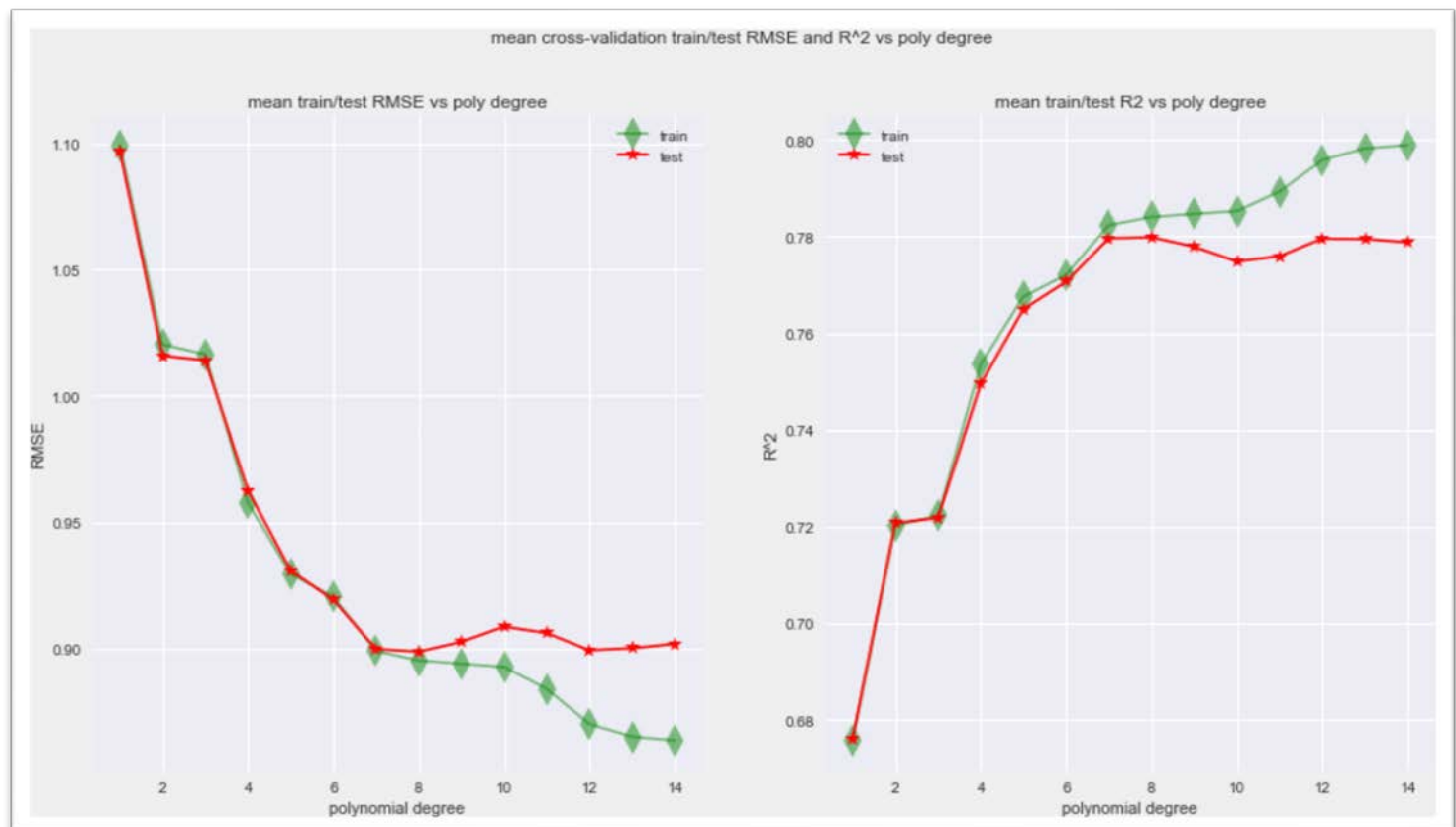
Many modeling techniques rely on minimizing the squared difference between predicted and actual values. Outlying values have very large squared error compared to inliers. As we see below, this accounts for relatively low metrics for models that seem to do a fairly good job of capturing the trend of the data.


### The Modeling Strategy and Approach

The approach taken was to work from simpler to more detailed models (e.g., less features to more features; higher bias to lower bias models). We used multiple kinds of techniques to be able compare results.

For every model, we made a graphical; comparison of fitted and actual values to aid in assessment of the mode (see Figures 15-23).

FIGURE 14. GRID SEARCH CROSS VALIDATION OUTPUT





For each model, if necessary, we analyzed hyper-parameter values for goodness of fit (bias/variance), choosing hyper-parameters with best characteristics, rejecting overfit models. To tune hyper-parameters, we used grid-search with 3-fold cross-validation (see Figure 14)

Then the best cross-validated model was applied to the test set for performance measurement.

Where appropriate, we added regularization to evaluate its utility.

## The Models

The following models were fitted and tuned for best performance:

- **Model 01:** Baseline OLS linear model: cost ~ time: test set  $R^2$ : 0.4043
- **Model 02:** OLS linear model: cost ~ time, size, state: test set  $R^2$ : 0.4613
- **Model 03:** OLS polynomial model (degree=14): cost ~ time, size, state: test set  $R^2$ : 0.520
- Regularized versions of OLS polynomial models (Ridge/Lasso) with best parameters yielded no improvement in performance.
- **Model 04:** Random Forest Regressor with best parameters cost ~ time, size, state: test set  $R^2$ : 0.5993
- **Model 05:** OLS linear model: median(cost) ~ time, weeks: test set  $R^2$ : 0.715
- **Model 06:** OLS polynomial model (degree=8): median(cost) ~ time (weeks): test set  $R^2$ : 0.8341
- **Model 07:** OLS polynomial model (degree=8): median(cost) ~ time(months), size: test set  $R^2$ : 0.8512
- **Model 08:** OLS polynomial model (degree=8): median(cost) ~ time(months), size, state: test set  $R^2$ : 0.75

We began with an Ordinary Least Squares linear model with time as the single predictor to capture a baseline model and get a sense of the complexity in the data. This simple model captures about 40% of the variance in the data.

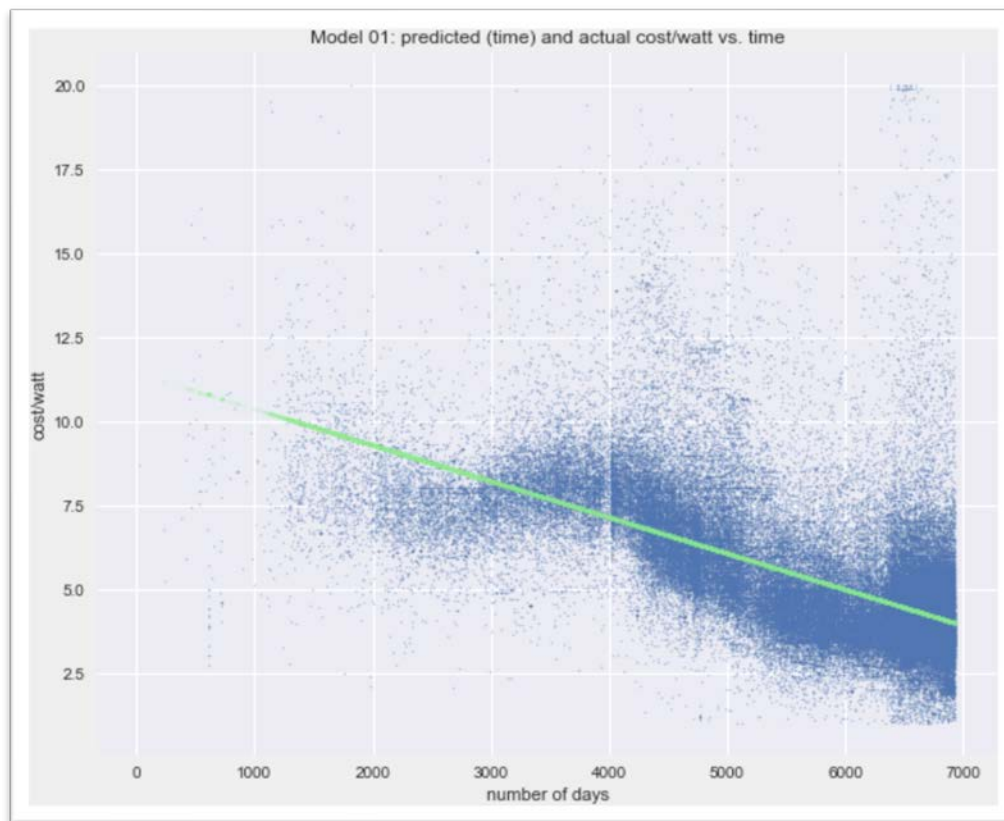


FIGURE 15. MODEL 01: BASELINE OLS LINEAR MODEL:  $\text{COST} \sim \text{TIME}$ :  $R^2: 0.4043$

We next add size (kw) and state as predictors along with time. Adding size and state to the linear model brought the R2 up to 0.46.

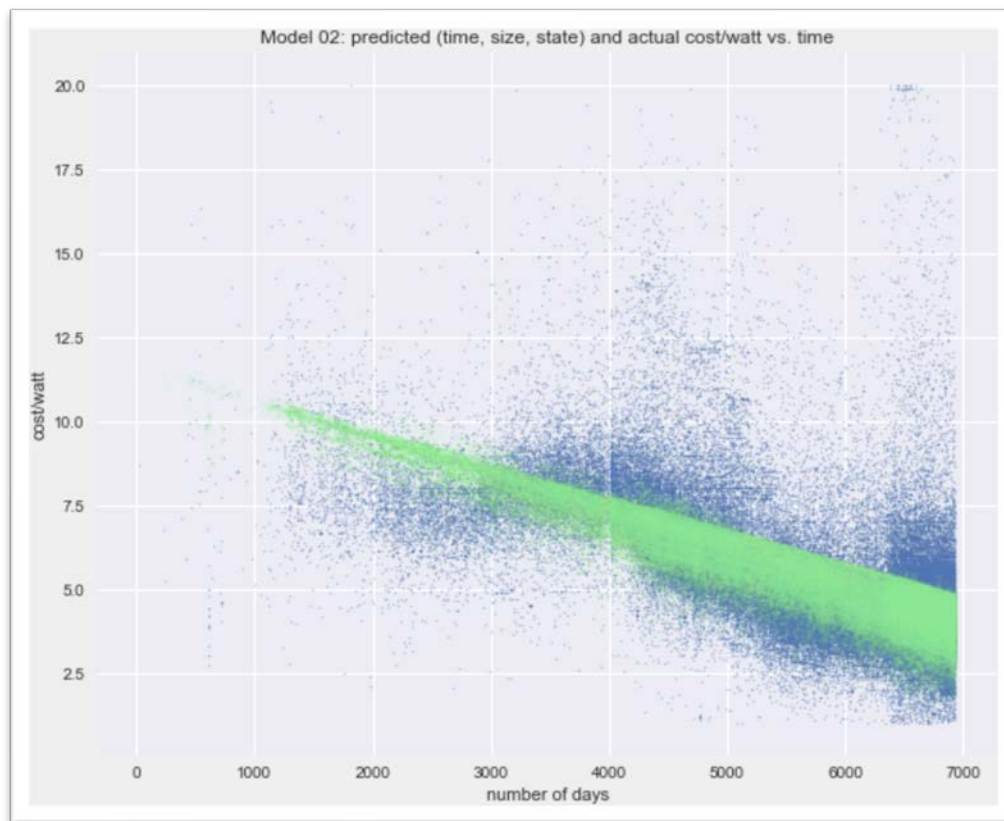


FIGURE 16. MODEL 02: OLS LINEAR MODEL:  $\text{COST} \sim \text{TIME, SIZE, STATE}$

We then introduce a polynomial expansion to the OLS framework, incrementally increasing the polynomial degree, stopping when the model is overfit, as indicated by a drop in R2 on the test set. The best generalizable fit to the data was polynomial degree 14. This much more flexible model has R2 of 0.53.

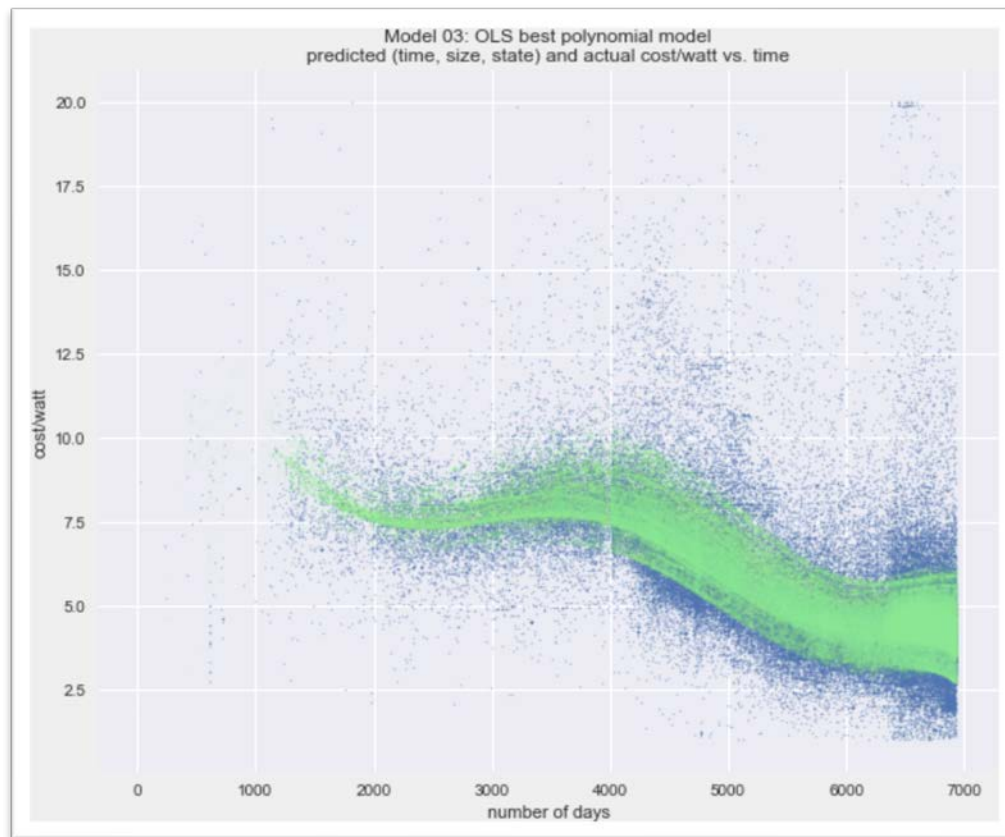


FIGURE 17. MODEL 03: OLS POLYNOMIAL MODEL:  $\text{COST} \sim \text{TIME, SIZE, STATE}$

At this point, we experimented with L1 and L2 regularization (Lasso and Ridge Regression) to see if we could achieve better fit and generalization, varying both polynomial degree and the regularization parameter. The best results were no better than the unregularized model.

Having gone as far as possible down this OLS-based path, we built a Random Forest model, a non-parametric model. The hyper-parameters number of estimators and tree depth were explored with grid-search and cross-validation. The model with best parameters had R2 of 0.60 on the test set, a decidedly stronger value than OLS-based models.

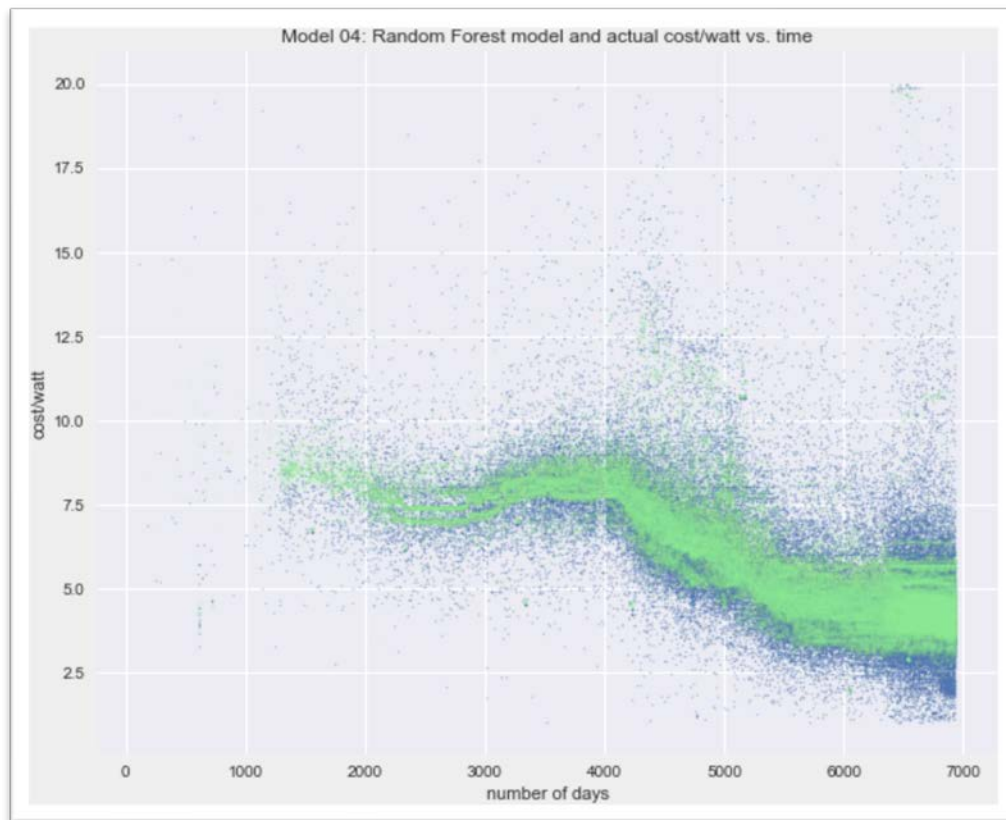


FIGURE 18. MODEL 04: RANDOM FOREST REGRESSOR WITH BEST PARAMETERS  $\text{COST} \sim \text{TIME, SIZE, STATE}$

The Random Forest Regressor achieved a better  $R^2$  by capturing some characteristics of the noise. This is visible in the graphic where outlying predictions are easily seen.

This resulted in a key insight: to answer the original business question, we do not need a perfect model of solar pricing, we need a model that can be used to predict a Fair Model Value for solar installation over the next 12-24 months. A fair market value for cost is better represented by an average or median cost over a time window.

This suggested another approach, modeling of an aggregate of the price within time windows. Both the mean and the median provide noise rejection by reducing many prices to a single number. The random variations within the groups tend to cancel out yielding a less noisy data set. The mean is, however, sensitive to outliers (there are many in the dataset). The median rejects more noise since the amplitude of an extreme value is ignored.



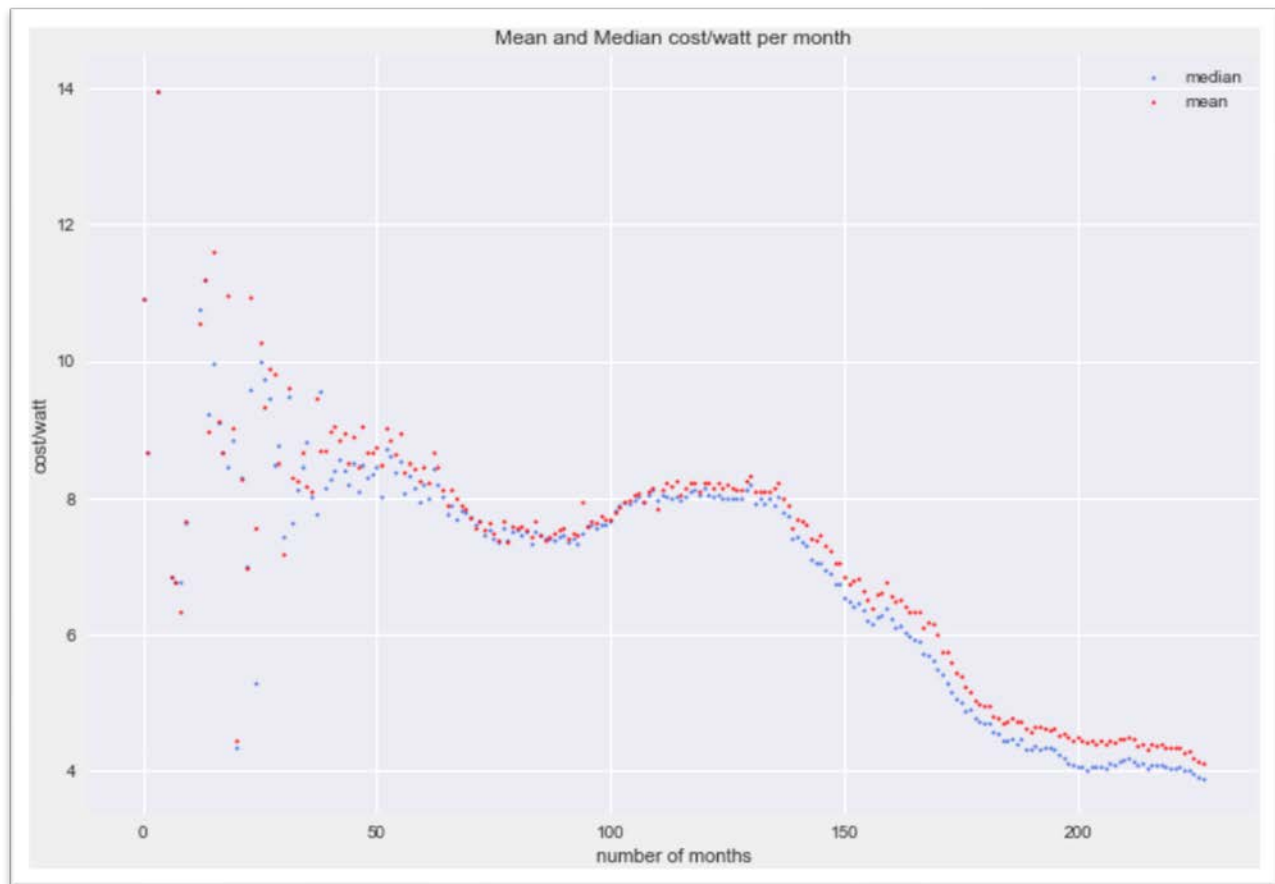


FIGURE 19. COMPARISON OF MEAN AND MEDIAN COST

The figure above shows a comparison between the mean and median of the data, where median is taken every month. The median is substantially below the mean because of the large number of outlying data at high prices. The median is a better representation of the Fair Market Value since half the prices are higher and half are lower, while the mean is inflated by many extreme positive values.

We began the new approach by using an OLS linear model of the median cost for comparison with the first model. We examined the median taken at different time windows (day, week, month). The best performing time frame for median was weeks with R2 of 0.715 (as compared to 0.401 for the direct linear model, Model 01).

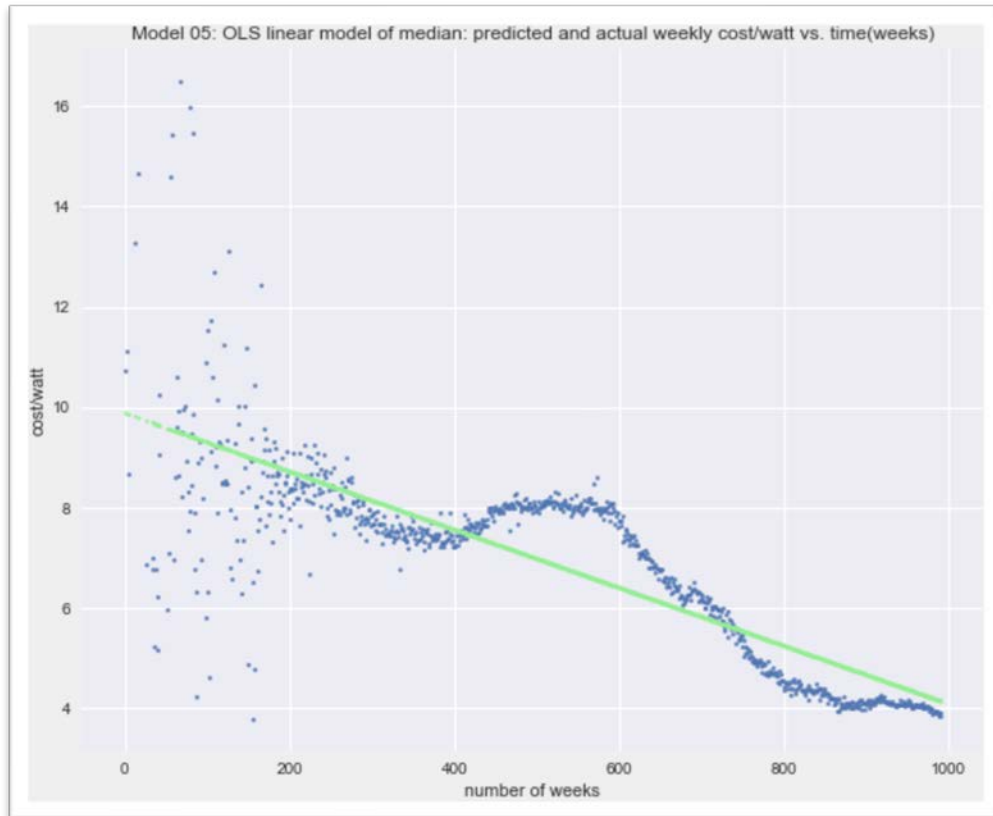


FIGURE 20. MODEL 05: OLS LINEAR MODEL:  $\text{MEDIAN}(\text{COST}) \sim \text{TIME}(\text{WEEKS})$

We continued with polynomial expansion of the median cost at different time windows (day, week, month) in Model 06. The best performing time frame for median was weeks with R2 of 0.834 at polynomial degree 8 (compare to Model 03, R2: 0.520).

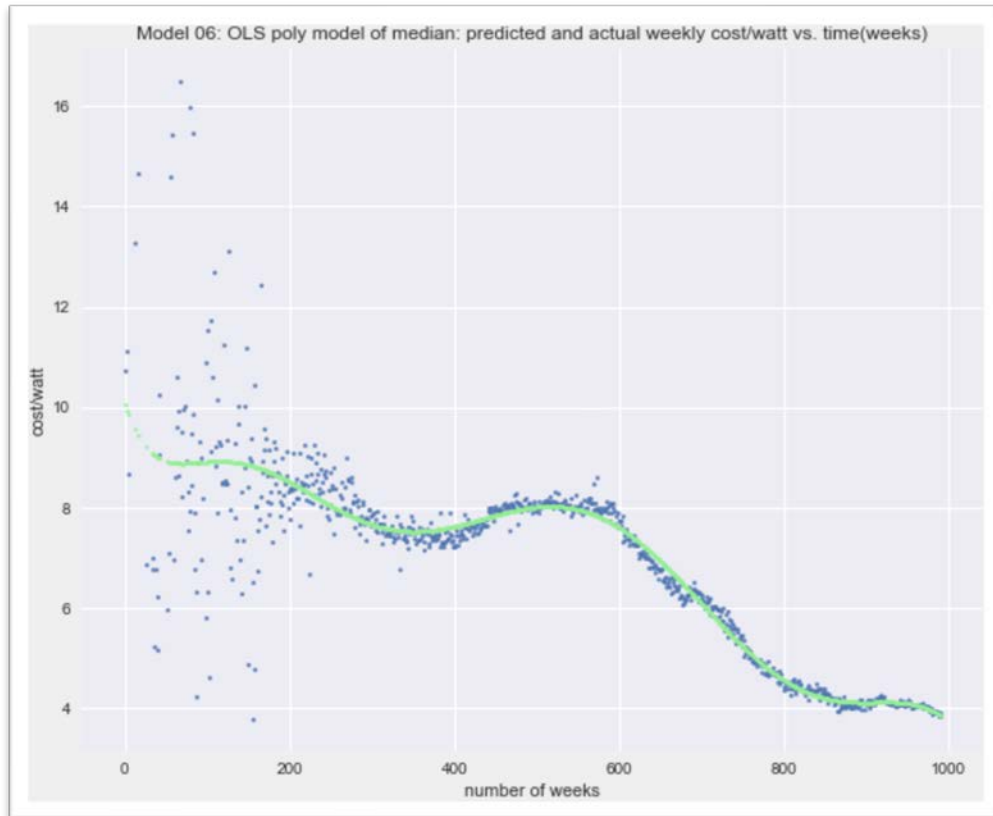


FIGURE 21. MODEL 06: OLS POLYNOMIAL MODEL: MEDIAN(COST) ~ TIME (WEEKS)

We then added size to the model by binning the size data and taking the median over combination of time interval and size. This increased R2 to 0.851.

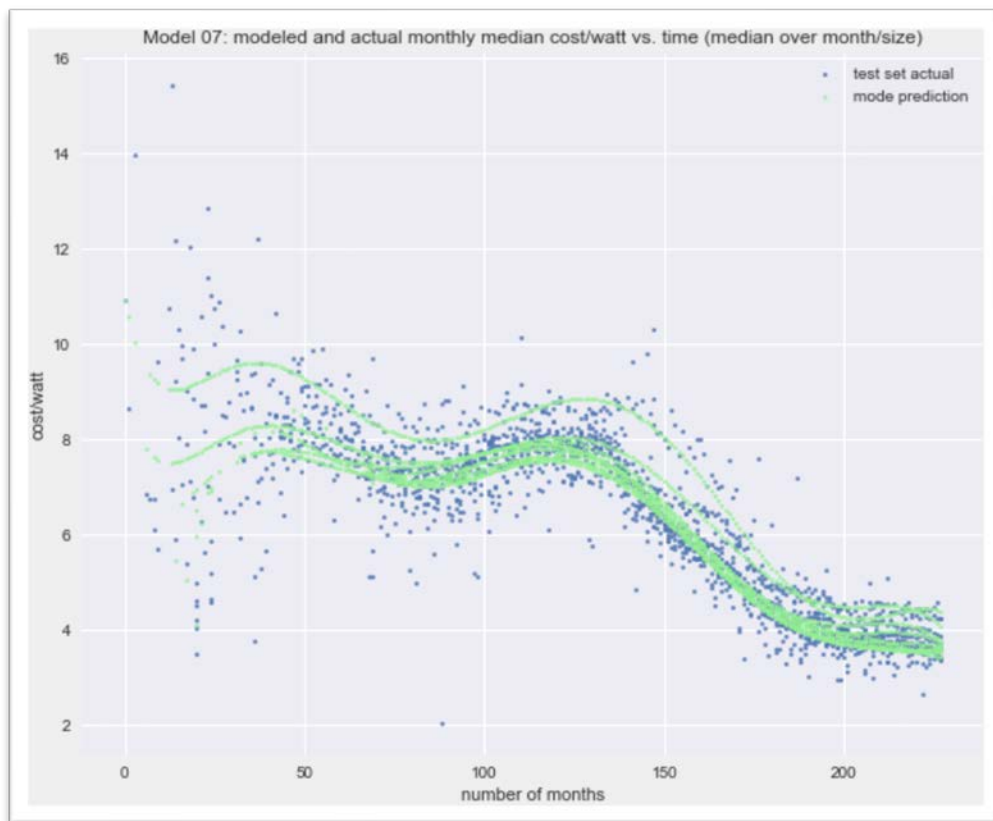


FIGURE 22. MODEL 07: OLS POLYNOMIAL MODEL:  $\text{MEDIAN}(\text{COST}) \sim \text{TIME}(\text{MONTHS}), \text{SIZE}$

For the next model, we added state to previous model (Model 08). This did not improve performance, in fact reducing the best R2 to 0.75 (though still comparing favorably with 0.46, the corresponding value for Model 03). This reduction may be due to poor grouping in the median (now over three features, time, size group and state). As we saw in Figure 6 above, five states constitute the bulk of the US market. The aggregation over small state/size groups may reduce the signal/noise ratio, compared to the previous model.

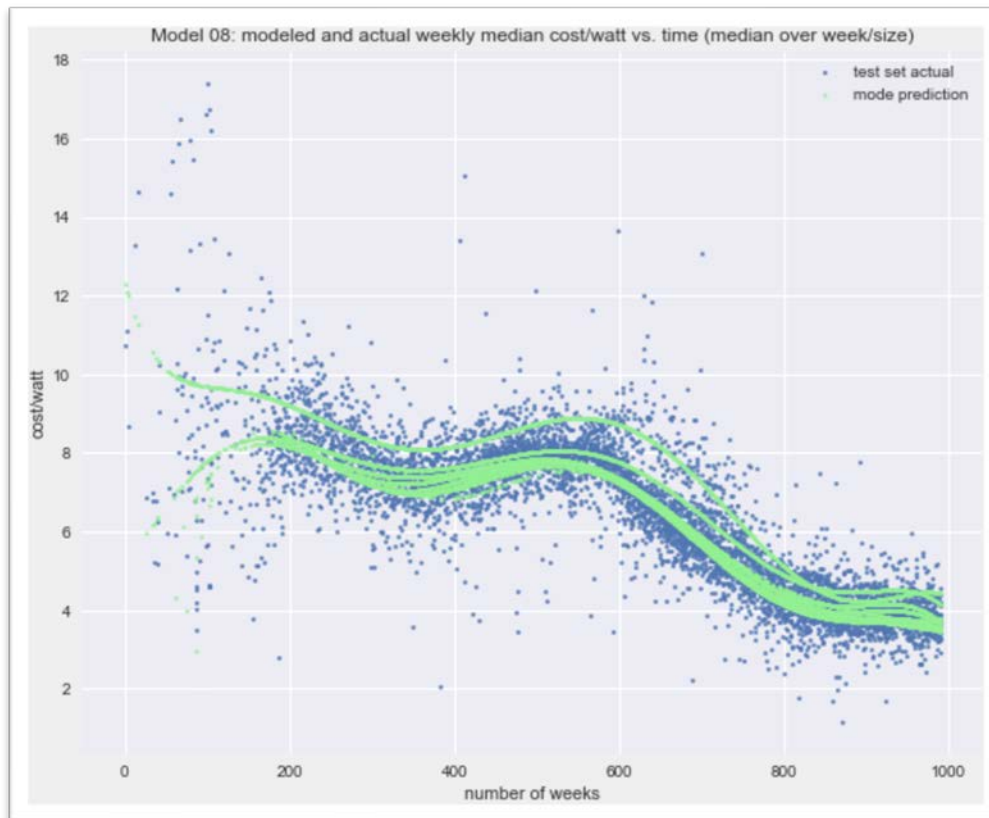


FIGURE 23. MODEL 08: OLS POLYNOMIAL MODEL:  $\text{MEDIAN}(\text{COST}) \sim \text{TIME}(\text{MONTHS}), \text{STATE}$

## Predictions

While a strong R2 is an indication of a good fit to the dataset, for our purposes, we require a model that provides reasonable extrapolations for the 18-24 months past the end of the data since the shape of the extrapolation curve is fundamental to the business question at hand. If the curve is flat or sloping upwards it is likely that there are no or little savings to be accrued by deferring the purchase of a solar installation. If the slope is moderately or steeply negative, waiting to purchase is likely to be rewarded with substantially lower cost.

As a result, we need not only an accurate fit (providing a good starting point for extrapolation), but also stable predictions, within market constraints. The market constraints are two-fold:

- Price is unlikely to increase substantially (dropping materials costs, vendor competition, continued government incentives, etc.)
- Installation cost provides a positive floor for cost.

Below we compare each of the candidate predictions graphically.

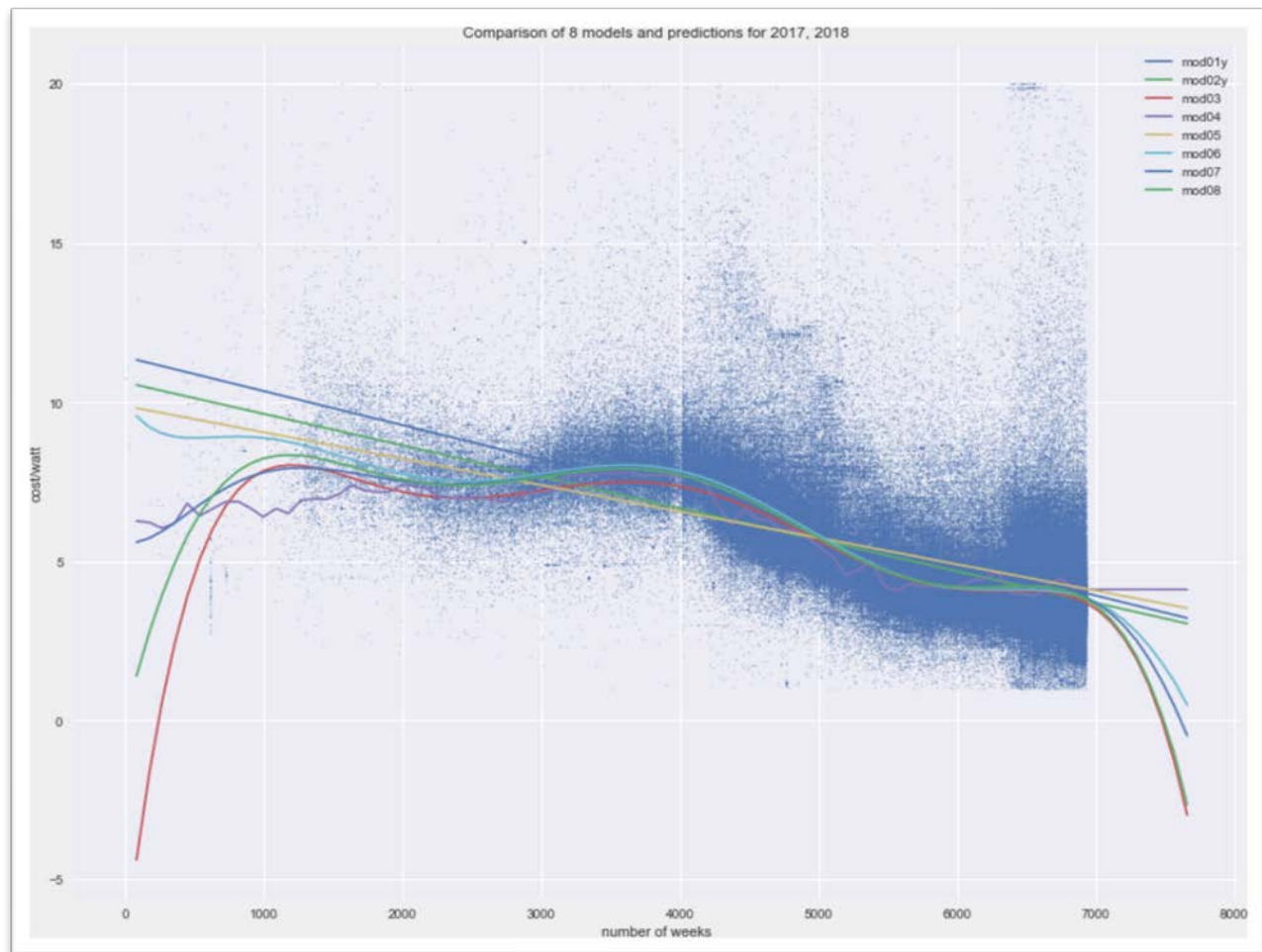


FIGURE 14. COMPARISON OF 8 MODELS

The less flexible linear models seem to underestimate the rate of decline in cost, while the highly flexible OLS-based models seem generally to overestimate the decline (in two cases resulting in negative cost in 2018)

Model 06 captures the trend of the data while remaining positive throughout 2017 and 2018.

The Random Forest Regressor predicts a constant cost in 2017 and 2018.

The graph below depicts the average of all eight models compared to Model 6. This combined prediction has both the stability of lower-bias models and high quality fit associated with the higher bias (and more accurate) models

We choose this combined model as our prediction for the upper and lower bounds of cost/watt in 6-8 quarters after the end of the data (2017-12-31).

The numerical forecasts are shown in Table 3.

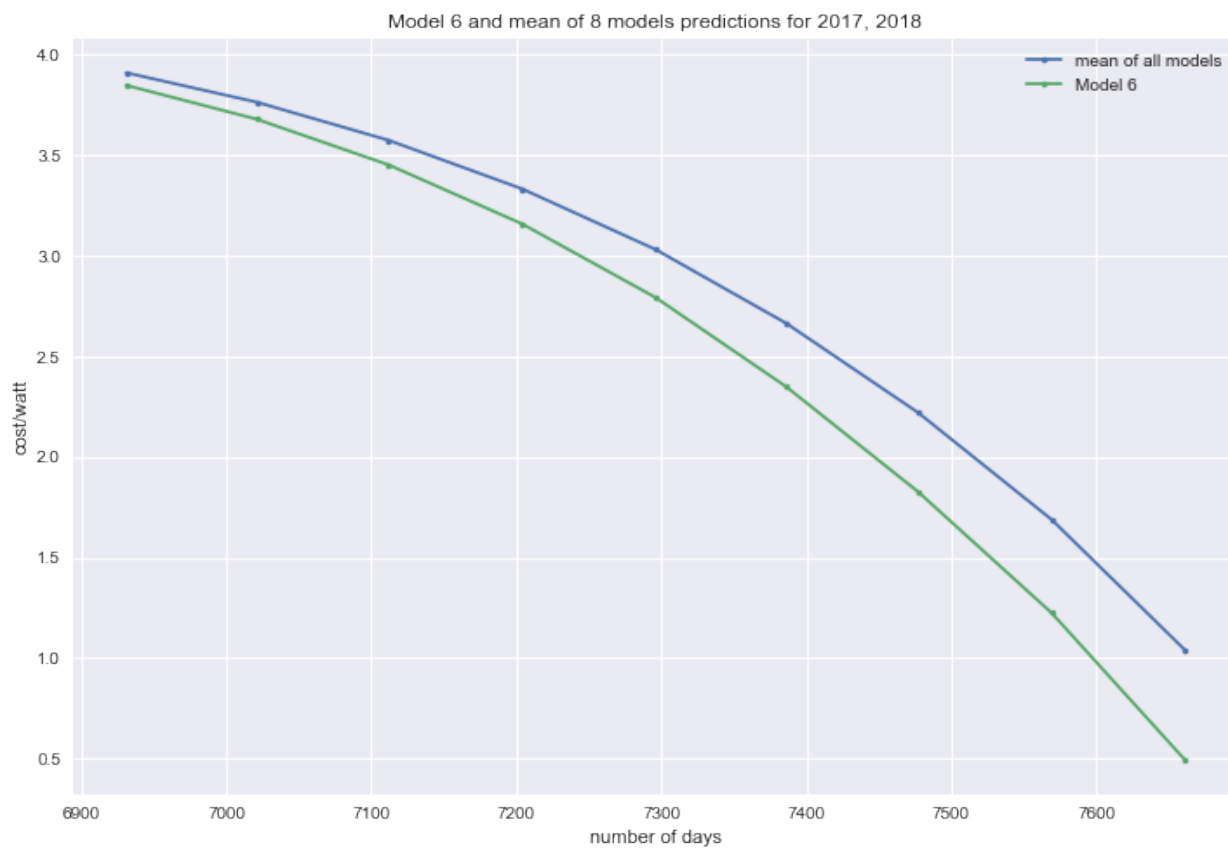


FIGURE 25. PREDICTED UPPER AND LOWER BOUNDS FOR COST/WATT IN 2017 AND 2018

TABLE 5. PREDICTION OF UPPER/LOWER BOUNDS FOR 8 QUARTERS AFTER END OF DATA

	20161231	20170331	20170630	20170930	20171231	20180331	20180630	20180930	20181231
Mean prediction	3.91	3.76	3.57	3.33	3.03	2.66	2.22	1.69	1.04
Model 6 Prediction	3.85	3.68	3.45	3.16	2.79	2.35	1.83	1.22	0.49



## Recommendations

The driving question for the product is: “Will it be more cost-effective to install now, or will I save money by waiting a year or two?”.

Having explored and modeled the data, we are now able to make well-informed recommendations.

- **Get several competitive quotations for the solar installation.**

The data show that prices for similar solar installation vary dramatically (in some case by a factor of 10). A selection of vendors is likely to provide the best price. Getting a competitive price is probably more important than optimal timing of the purchase.

- **Given the shape of the predicted cost curve, it is reasonable to expect that median cost will drop by about \$1.00/watt over 2017.**

It is also reasonable to expect at least this rate of decline in 2018 and there are grounds to believe that the rate of price decline will accelerate in 2018. It would be prudent to revisit this analysis incorporating more data as it becomes available. Given these predictions the customer can weigh the current cost of installation and the expected price decrease and make an informed decision.

## Future work

The techniques used in this analysis are by no means the only modeling tools available.

There are many other regressors available including RANSAC, Theil Sen and HuberRegressor, which all have a degree of robustness with respect to outliers.

There are also a set of techniques specifically tailored to analysis and prediction of time series data. These methods typically use some abstraction of past time values as an input for prediction. These tools may provide more ways to tailor a regressor more suitable for extrapolation.

Within the framework given here, other transformations of the predictor data are worth considering. Specifically, a logistic function (akin to  $f(x) = C(1 - 1/(1 + e^{kt}))$ ) may provide an workable alternative to polynomial transformation. A dimensional reduction, particularly with regard to the state data may improve model performance.

A future project could explore these tools in this setting. The current work would then provide a well-founded basis for comparison.