­

**Features Impacting Residential Solar Project Costs in San Diego, California**

**Executive Summary**

Christopher Fischer - 011933891

College of Information Technology, Western Governors University

Dr. Daniel Smith

December 19, 2024

Features Impacting Residential Solar Project Costs in San Diego, California

Executive Summary

# Problem Statement

San Diego, California, boasts one of the highest solar adoption rates in the United States. However, to reach the city’s goals of net-zero emission of greenhouse gases by 2035, the rate of adoption will need to accelerate (City of San Diego, 2024). One consistently reported barrier hindering residents from choosing to go solar is the cost of installing a system (National Renewable Energy Laboratory, 2016). If a reliable cost-prediction model can be built, a customer-facing tool could be offered to allow a homeowner to quickly receive a rough estimate after answering just a few questions. Such a tool could help lower reluctance to begin the process (Tong, 2016). The first step in producing such a model is to identify features of a solar project that have a statistically meaningful impact on the target variable.

With this context in view, this study attempts to answer the question: *To what extent can the features of a proposed residential solar project (system size, single/multi-phase, self-installed, number of modules, module capacity, module efficiency, and battery capacity) affect its total installed cost?* The study’s hypotheses are:

**H0** – The selected features do not statistically impact the total project cost.

**Ha** – The selected features statistically impact the total project cost.

# Data Analysis Process

The data used for this study come from the “Tracking the Sun” project run by the Lawrence Berkeley National Laboratory (<https://bit.ly/trackingthesun-2024>). The dataset contains statistics on over 3.7 million solar projects installed across 31 states through 2023 (Electricity and Markets Policy Group, 2024).

As this study aimed to improve solar adoption rates in the greater San Diego area, the dataset was reduced to include only residential solar projects installed in that region during 2023. The data was anonymized by setting all city values to “San Diego” and all zip code values to “92101”. The resulting dataset consists of 43,265 observations.

The tools used for this study included Python, Jupyter Notebook, and Visual Studio Code. The following Python libraries were used to support this study:

|  |  |
| --- | --- |
| Module | Purpose in this Study |
| Pandas | Provides tools for managing datasets |
| Matplotlib.pyplot | Offers tools for visualizing data and results |
| Seaborn | Additional visualization capabilities |
| Shap | Visualizes feature contributions to the results of a model |
| Statsmodels | Provides a set of tools for creating and evaluating models, particularly well-suited for statistical analysis and model interpretation (Srinivasan, 2020) |
| Sklearn | Implements scaling and additional regression methods |

Table 1 - Python Libraries Supporting Study

The data were prepared for study as follows:

1. Evaluate and clean missing data
   * 828 observations with missing data were removed
2. Transform certain features to appropriate data types
   * Four features were transformed from float to integer
3. Identify and treat potential outliers – observations with outlying values were removed from further study
   * Outliers were identified in *PV\_System\_Size\_DC, module\_quantity\_1, nameplate\_capacity\_module\_1, efficiency\_module\_1*, and *battery\_rated\_capacity\_kW*
   * All *battery\_rated\_capacity\_kW* outliers were determined to be acceptable and were retained
   * 1,393 observations with other outlying values were removed
4. Perform univariate and bivariate exploratory data analysis, including correlation
   * Several features demonstrated correlation to the target variable

A screenshot of a graph

Description automatically generated

1. Reduce dimensionality through feature elimination
   * Variable Influence Factors (VIF) were used to eliminate *module\_quantity\_1* to remove excessive multicollinearity
2. Normalize feature values through scaling

Once the data were prepared, an Ordinary Least Squares (OLS) multiple linear regression model was fitted. Since OLS regression assumes homoscedasticity, a Breusch-Pagan test was performed on the resulting model (Frost, 2019). The results indicate that no heteroscedasticity was observed.

# Findings

The resulting model yielded an R2 measure of 0.52. This indicates the model is a moderate fit, doing a reasonable job of explaining the data (52% of the variation in the response variable) but leaving a significant portion of variability unexplained. Nevertheless, the coefficients table reveals that each of the retained predictor variables contributes in a statistically significant degree to the model’s results, as shown in the following table.

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Coefficient** | **P-value** |
| **PV\_system\_size\_DC** | 0.6525 | 0.000 |
| **battery\_rated\_capacity\_kW** | 0.2165 | 0.000 |
| **self installed** | -0.1023 | 0.000 |
| **efficiency\_module 1** | 0.0586 | 0.000 |
| **nameplate\_capacity\_module 1** | -0.0096 | 0.006 |
| **multiple\_phase\_system** | -0.0088 | 0.011 |

Table 2 - Predictor Variable Analysis Grouped by Contribution Strength

All predictors demonstrate statistical significance by having very low p-values. However, the coefficients show that the magnitude of each variable’s contribution varies widely. The predictors are arranged in three groups: those with a high degree of contribution, those with a moderate degree, and those with minimal contribution.

The SHAP graph below aligns nicely with the team’s analysis of the variable coefficients table. At a glance, viewers can grasp the importance of each predictor in the response variable for each observation.

A graph of a diagram

Description automatically generated with medium confidence

In summary, the results of this study show that the null hypothesis may be rejected. Several selected features statistically impact the total installed cost of residential solar projects in San Diego, California.

# Limitations

One limitation of this study is that no variables that characterize the building upon which the solar project was installed were available. Such variables might include structure square footage, number of floors, and the roof's size, shape, slope, and azimuth. (Office of Energy Efficiency & Renewable Energy, n.d.) These variables likely influence the cost of a solar project.

Another limitation is that OLS linear regression assumes a linear relationship between the predictor and response variables. The further from this ideal the data are, the less reliable the results of the regression will be. The predictors in this study show a *generally* linear relationship to the target, but not perfectly so.

# Recommendations

In response to the first limitation identified above, the team recommends that stakeholders seek additional data sources to augment the inputs to the linear regression model. Doing so could provide additional statistically relevant variables to help build a model for predicting solar project cost.

The team also suggests two additional avenues of study that may benefit stakeholders. First, while this study examined the total project cost, another interesting way of assessing a cost is to determine the dollars per kilowatt generated for the project. This would allow analysts to discover variables that may improve a project's economic efficiency. Second, the study could be widened to include more geographic diversity of observations. This study focused on one city in California. However, there may be broad benefits to having a robust regional or even national solar project cost-predictive model.

# Expected Benefits

This study is a valuable proof-of-concept exercise demonstrating the feasibility of developing a model for predicting the expected cost of a residential solar power project from several inputs. While further work is required to build such a model, knowing that there is verifiable statistical relevance to a set of predictor variables will give model-builders a significant head-start in their efforts.

# References

City of San Diego. (2024). *Climate Action Plan 2023 Annual Report* (pp. 5, 8–10). <https://api.watch.kausal.tech/documents/446/2023_CAP_Annual_Report_9-23-24_AM_FINAL.pdf>

Electricity and Markets Policy Group. (2024, August 21). *Tracking the Sun | Electricity Markets and Policy Group*. Lawrence Berkeley National Laboratory. <https://emp.lbl.gov/tracking-the-sun>

Frost, J. (2019, March 15). *Heteroscedasticity in Regression Analysis*. Statistics by Jim. <https://statisticsbyjim.com/regression/heteroscedasticity-regression/>

National Renewable Energy Laboratory. (2016). *Solar Energy Evolution and Diffusion Studies: 2014–2016 Customer Motivations and Adoption Barriers Study*. Nrel.gov. <https://www.nrel.gov/solar/market-research-analysis/2014-2016-study.html>

Office of Energy Efficiency & Renewable Energy. (n.d.). *Homeowner’s Guide to Going Solar*. Energy.gov. Retrieved December 14, 2024, from <https://www.energy.gov/eere/solar/homeowners-guide-going-solar>

Tong, J. (2016, June 20). *How to Turn Solar “Considerers” Into Solar Adopters*. Greentechmedia.com; Greentech Media. <https://www.greentechmedia.com/articles/read/How-to-Turn-Solar-Considerers-Into-Solar-Adopters>