EPIC Final Report

Program	**Electric Program Investment Charge (EPIC)**
Administrator	**San Diego Gas & Electric Company**
Project Number	**EPIC-2, Project 6**
Project Name	**Collaborative Programs in RD&D Consortia**
Module Name	**Demonstration of Methodology and Tools for Estimating Propensity for Customer Adoption of Photovoltaics**
Date	**December 31, 2017**

Project Attributions

This comprehensive final report documents the work done in one module of EPIC-2, Project 6.

The project team that contributed to the project definition, execution, and reporting included the following individuals, listed alphabetically by last name.

SDG&E

Bui, Nancy

Goodman, Frank

Katmale, Hilal

Mazhari, Iman

Salmani, Amin

Wilson, Dan

Navigant Consulting/Trove

Corfee, Karin

Gifford, Will (Trove)

Goffri, Shalom

Romano, Andrea

Seiden, Ken

Welch, Cory

Executive Summary

Overview

The objective of Electric Program Investment Charge 2 (EPIC-2), Project 6 (Collaborative Programs in RD&D Consortia) is to accomplish highly leveraged demonstration work through industry collaborative R&D organizations. The focus of this project module was to identify methodologies and tools for determining the primary drivers for residential photovoltaic (PV) adoption, predict residential PV adoption over time, and to demonstrate selected methods on a use case (e.g., propensity to adopt PV on the ZIP code level). The effort also developed recommendations about whether to adopt all or some of the methods and tools on a commercial basis. The project team focused specifically on residential sector PV market adoption. Additionally, the project team conducted machine learning (ML) analytics on disadvantaged communities (DAC) ZIP codes and evaluated the difference in propensity to adopt solar PV between DAC and other ZIP codes.

The scope of this project demonstrated methodologies and tools for forecasting the propensity for residential customer solar PV adoption in California and SDG&E ZIP Codes. The project included the following major tasks.

- Literature Review and Methodology Justification
- Methodology Framework Development
- Demonstration Plan
- Disadvantaged Communities Analysis

Using ML, the project team identified the most important attributes driving adoption at the non-DAC and DAC ZIP code level as detailed in **Error! Reference source not found.**.

Table E- 1 Machine Learning Key Customer Attributes

Non- DAC Key Customer Attributes	DAC Key Customer Attributes
• Average number of credit mortgage type inquiries in the last 12 months	• Average number of credit mortgage type inquiries in the last 12 months
• Climate zone	• % Owner Occupied Households
• Percentage of households that are married families	• Climate zone
• Average balance on an open auto loan and lease trades reported in the last 6 months	• % Manufactured Housing Units
• Average household size	
• Percentage of owner-occupied housing units	

Comparing adoption in DAC and non-DAC ZIP codes, owner occupancy emerged as a key attribute explaining the difference in PV market share. The percentage of owner occupied homes is 63% for non-DAC ZIP codes, compared to 50% for DAC ZIP codes.

Key Findings and Conclusions

The strength of aggregate and ZIP code-level back-casts suggest that causal models can be used to forecast residential rooftop PV adoption moving forward with a reasonable degree of accuracy, even when

the analysis is spatially disaggregated. Such methods could support integrated resource planning and a better understanding of likely solar PV installation location. As demonstrated by the results, the key findings of this project can be summarized as follows:

- Causal models, when appropriately calibrated, can explain historical adoption patterns well.
- Preliminary results suggest that the adoption of residential solar PV in California and in the SDG&E service territory is past the inflection point in the characteristic S-curve of adoption.
- ML techniques can help to explain historical adoption patterns and can reduce the variance between simulated and actual adoption when analyzed at a granular level.
- The rate of diffusion (i.e., the shape of the S-curve) is reasonably consistent spatially; the long-run market share appears to differ more substantially when analyzed at a granular level (e.g., at the ZIP code level).
- ZIP code-level forecasts seem possible with reasonable accuracy.
- Market share solar PV as a percent of total homes is about 50% higher in non-DAC ZIP codes (6% vs. 4% market share).
- Owner occupancy is a key attribute in explaining the differences between DAC and non-DAC solar PV market adoption.
- Statewide analysis of Non-DAC and DAC ZIP codes generates close fit between simulated and actual.

Recommendations

The project team recommends that SDG&E not commercially adopt these methods and tools at this juncture, without more foundational work being done first. Based on the pre-commercial demonstration results and findings in this project, the following actions by SDG&E or other stakeholders are recommended as steps toward commercial adoption of the demonstrated methods and tools.

- Improve SDG&E's existing zip-code based Bass diffusion technique with refinements for the long-run market share parameters based on significant customer attributes.
- Improve certain model inputs (e.g., historical PPA prices, kilowatt-hour production, technical suitability due to shading and orientation, price sensitivity, and correlation between homeownership and credit scores).
- Leverage the same or equivalent methodology to evaluate solar PV adoption for other specific segments of interest and potentially individual customer analysis, including but not limited to: commercial and industrial customers, low-income customers, and customers on distribution feeders that are capacity constrained or at risk for reverse power flow.
- Adapt the methodology for use in forecasting adoption of other DER types.
- Consider utilizing a customer discrete choice survey approach to facilitate independent estimation of both the long-run market share parameters and the Bass diffusion coefficients.

Table of Contents

Executive Summary ..... **iii**

Overview iii

Key Findings and Conclusions iii

Recommendations iv

1. Introduction **1**

1.1 Project Objective 1

1.2 Project Approach 1

1.2.1 Task 1: Formation of the internal SDG&E project team 1

1.2.2 Task 2: Development of a project plan and contractor selection 1

1.2.1 Task 3: Demonstration Activities 2

1.2.2 Subtask 3.1: Literature Review and Methodology Justification 2

1.2.3 Subtask 3.2: Methodology Framework Development 9

1.2.4 Subtask 3.3: Demonstration 11

1.2.5 Subtask 3.4: Disadvantaged Communities Analysis 17

1.2.6 Task 4 – Conduct Data Analysis and Develop Findings and Recommendations 18

1.2.7 Task 5 – Final Report 18

2. Results and Analysis **19**

2.1 Demonstration Activity Results and Analysis 19

2.1.1 Customer Attributes Driving Adoption 19

2.1.2 Data Compilation 19

2.1.3 Random Forest Model 21

2.1.4 Granular Adoption Analysis 23

2.2 Disadvantaged Communities Results and Analysis 32

2.2.1 Customer Attributes Driving Adoption 32

2.2.2 Data Compilation 32

2.2.3 Machine Learning Models 33

2.2.4 Statistical Analysis Results 35

2.2.5 DAC vs. Non-DAC Adoption Analysis 37

2.3 Assumptions 38
3. Key Findings 39
4. Recommendations and Next Steps 41
5. Metrics and Value Proposition 42
5.1 Metrics 42
5.2 Value Proposition 42
6. Technology Transfer Plan 43
6.1 SDG&E Technology Transfer Plans 43
6.2 Adaptability to Other Utilities and Industry 43
7. References 44

Figures

Figure 1. Task 3 Demonstration Activities Summary 2
Figure 2. Adoption Forecasting Methodology for SDG&E 10
Figure 3. Ensemble Model: Example for Regression 13
Figure 4. Illustrating Difference Between Rate of Diffusion and Long-Run Market Share 14
Figure 5. Illustration of a Calibration of Bass Diffusion Coefficients – Comparison of Actual Historical vs. Model Simulated Adoption 15
Figure 6. Logit Market Share Model 16
Figure 7. NEM Installation Percentage by ZIP Code 20
Figure 8. Incremental and Cumulative MW Installed in Modeled ZIP Codes – Demo Results 21
Figure 9. Random Forest Estimation Error Distribution for Percentage of Households with NEM Installations 22
Figure 10. Variable Importance Plot – Estimating Percentage of NEM Installation Households 23
Figure 11. Model Simulated vs. Actual PV Installed – Statewide Coefficients Only 24
Figure 12. Simulated Cumulative Adoption (2009-2017) for all 118 SDG&E ZIP Codes – Statewide Coefficients Only (Held Constant Across All ZIP Codes) 25
Figure 13. Model Simulated vs. Actual PV Installed – Adding Customer Attribute Coefficients 27
Figure 14. Simulated Actual Cumulative Adoption (March 2017) for all 118 SDG&E ZIP Codes – Adding Customer Attribute Coefficients 27
Figure 15. Model Simulated vs. Actual PV Installed 29
Figure 16. Simulated vs. Actual Cumulative Adoption (in March 2017) for all 118 SDG&E ZIP Codes – Adding ZIP Code-Specific Coefficients 29
Figure 17. Monthly Simulated vs. Actual Cumulative Adoption (January 2009 – March 2017) for all 118 SDG&E ZIP Codes – Adding ZIP Code-Specific Coefficients 30
Figure 18. Simulated vs. Actual Cumulative Adoption over Time – ZIP Codes 92020, 91977, 91910, and 91906 31
Figure 19. Simulated vs. Actual Monthly Incremental Adoption over Time – ZIP Codes 92020, 91977, 91910, and 91906 32
Figure 20. NEM Installation Percentage, by DAC ZIP Codes 33
Figure 21. Random Forest Attribute Variable Importance – DAC ZIP Codes 34
Figure 22. Random Forest Estimation Error Distribution for Percentage of Households with NEM Installations 35
Figure 23. Percentage of Owner Occupied Homes and Market Share Attributes 36
Figure 24. Remaining Four Attributes 37
Figure 25. Comparison of DAC and non-DAC Adoption: Simulated vs. Actual 38

Tables

Table 1. BDCM Key Inputs and Outputs 10
Table 2. Public Data Sources Used for this Demonstration 12
Table 3 Additional Public Data Sources Used for DAC Demonstration 17
Table 4. Random Forest Estimation Error Quartiles and Mean Percentage of Households with NEM Installs 22
Table 5. Key Customer Attributes 23
Table 6. Random Forest Estimation Error Quartiles and Mean Percentage of Households with NEM Installs 35

List of Acronyms

BDCM	Bass Diffusion System Dynamics Causal Model
CART	Classification and Regression Tree
CTree	Conditional Tree Inference
DAC	Disadvantaged Communities
DER	Distributed Energy Resources
DG	Distributed Generation
DGStats	Distributed Generation Statistics
DR	Demand Response
EE	Energy Efficiency
EPIC	Electric Program Investment Charge
EV	Electric Vehicle
IOU	Investor-Owned Utility
kWh	Kilowatt-Hour
LCOE	Levelized Cost of Electricity
MASH	Multi-Family Affordable Solar Housing
ML	Machine Learning
MW	Megawatt
NEM	Net Energy Metering
NREL	National Renewable Energy Laboratory
PG&E	Pacific Gas & Electric
PPA	Power Purchase Agreement
PV	Photovoltaic
SASH	Single Family Affordable Solar Housing
SCE	Southern California Edison
SDG&E	San Diego Gas & Electric
T&D	Transmission and Distribution
TPO	Third-Party Owner
ZIP	Zone Improvement Plan

1. Introduction

1.1 Project Objective

The objective of EPIC-2, Project 6 (Collaborative Programs in RD&D Consortia) is to accomplish highly leveraged demonstration work through industry collaborative R&D organizations. The leveraging includes both prospective financial leveraging via co-sponsorship by other members of the collaborative, and intelligence leveraging by better informing the project content in EPIC activities with the knowledge of relevant activities occurring in a worldwide sense. The focus of this project module was to identify methodologies and tools for determining the primary drivers for residential photovoltaic (PV) adoption, predict residential PV adoption over time, and to demonstrate selected methods on a use case (e.g., propensity to adopt PV on the ZIP code [1] level). The effort also developed recommendations about whether or not to adopt all or some of the methods and tools on a commercial basis. The project team focused specifically on residential sector PV market adoption, envisioning that—depending on the degree of success in the demonstration—the methods and tools might someday be applicable to other areas, such as energy efficiency (EE), demand response (DR), non-PV distributed generation (DG), storage, electric vehicles (EVs), and microgrids. Additionally, the project team conducted machine learning analytics on Disadvantaged Communities (DAC) ZIP codes and evaluated the difference in propensity to adopt solar PV between DAC and other ZIP codes.

1.2 Project Approach

The project demonstrated tools and methodology for forecasting the propensity for customer adoption of DER in various parts of the SDG&E service territory. The major tasks were:

- Task 1: Formation of an internal SDG&E project team
- Task 2: Development of a project plan and contractor selection
- Task 3: Perform demonstration activities
- Task 4: Prepare comprehensive final report (draft for review and final version)
- Task 5: Final Report

1.2.1 Task 1: Formation of the internal SDG&E project team

Objective: Engage expertise needed to provide technical support.

Approach: An internal team consisting of engineers from the project's stakeholder groups within SDG&E was formed.

Output: A project team with structure and assignments.

1.2.2 Task 2: Development of a project plan and contractor selection

Objective – Prepare a detailed work plan, competitively procure a qualified contractor, and conduct a kickoff meeting with the selected contractor.

Approach – The internal team met with stakeholders within SDG&E and incorporated their inputs into the project plan. The project team carefully selected the data provided to the contractor, without disclosing proprietary and sensitive customer information. The team collaborated on finalization of the project plan.

Output – Procurement of contractor and completed project plan.

1.2.1 Task 3: Demonstration Activities

Figure 1 below describes the sub-tasks undertaken in Task 3 to demonstrate Methodology and Tools for Estimating Propensity for Customer Adoption of Photovoltaics.

Figure 1. Task 3 Demonstration Activities Summary

Based on the results of the demonstration, the project team identified challenges, recommendations, and next steps for future research.

1.2.2 Subtask 3.1: Literature Review and Methodology Justification

1.2.2.1 Objective

Conduct a literature review designed to provide a broad perspective on the methods, tools, and data necessary for distributed energy resource (DER) predictive analytics and adoption, with a specific focus on solar PV adoption by residential customers. The goal of this literature review was to provide both a theoretical and practical foundation to select a methodology, highlighting the benefits of machine learning-based enhancements to causal models.

1.2.2.2 Approach

The project team's literature review considered both academic and industry sources and focused on the following components:

- **Causal models**. Causal models are defined as the class of methods with closed functional forms/algebraic equations defining customer propensities and adoption forecasts. The project team has observed that the energy industry uses these types of models extensively for both EE and DER adoption. Examples include discrete choice models that provide a long-run probability of customer adoption (e.g., an equilibrium) and diffusion of innovation/bass models that explain the time dynamics and adoption path to that equilibrium.
- **Machine learning models**. Machine learning models are defined as a class of methods by which the functional form equations defining the relationships between input and output variables may not be known a priori and are instead learned through a training process. Multiple industries use these types of models to discover correlations between data and observed product adoption behaviors that might otherwise be missed.
- **Combined methods**. Both academic and industry researchers are working on techniques to combine causal and machine learning methods. Although this research remains in its infancy, early results suggest that causal models can be enhanced by data-driven, machine learning techniques.

1.2.2.2.1 Causal Models

The most common causal modeling paradigm for developing new technology adoption forecasts combines seminal theoretical constructs from the fields of economics and marketing science:

- **Consumer utility theory and discrete choice analysis.** Nobel Economics Laureate Daniel McFadden and others developed a class of consumer preference models fully consistent with rational consumers maximizing utility behavior over a discrete set of alternatives, and where the analysis outcome represents each alternative as probability of selection. Discrete choice analysis has been employed over the last 40 years to a variety of problems including transportation mode choice, energy forecasting and the choice of end-use systems and fuels, and existing and new product forecasting across a range of industries. As discussed further below, the key common feature of these models is that probabilities of selections are easily translated into market shares and product sales.

- **Diffusion of innovations and the Bass diffusion model.** Although discrete choice analysis provides a powerful theoretical and practical backbone for estimating long-run or equilibrium market shares of products and services, the modeling outcomes are largely time-invariant. Fortunately, Frank Bass and his colleagues in the marketing science field developed the Bass diffusion model to simulate the S-shaped approach to equilibrium that is commonly seen for technology adoption. In this model, market potential adopters flow to adopters through two primary mechanisms: adoption from external influences such as marketing and advertising, and adoption from internal influences, or word of mouth. The Bass diffusion model was first used to model the adoption of color TVs in the 1960s and has been used to forecast a broad variety of new technologies including computers, wireless telephones, smartphones, and now solar PV and other DER.

1.2.2.2.2 Consumer Utility Theory and Discrete Choice Analysis

Central to robust causal models is characterizing market share through established methodologies. This section summarizes key literature related to random utility theory and methods (such as discrete choice analysis) that can be used for market share parameter estimation.

- **D. McFadden. "Economic Choices," presented at the Prize Lecture, Stockholm, Sweden, Dec. 8, 2000. [2]**

  In his acceptance of the Nobel Prize in Economic Science, McFadden gave a lecture that discussed the "microeconomic analysis of choice behavior of consumers who face discrete economic alternatives." The theoretical basis for discrete choice analysis, McFadden describes the history, development, and application of the multinomial logit model, a means by which to determine the probability of choosing one alternative over others given their utility as characterized by their measurable attributes.

  The project team modeling approach employs the logit formulation developed and described by McFadden in its calculation of the long-run market share of solar PV. The construct is particularly suitable for this application, where the study endeavors to obtain more granular forecasts by identifying key customer attributes, which are readily incorporated into a logit market share formula.

- **M.E. Ben-Akiva, S.R. Lerman. Discrete Choice Analysis: Theory and Application to Travel Demand. Cambridge, MA: The MIT Press, 2006. [3]**

  In this work, the authors bring together the research of many to provide a comprehensive overview of discrete choice analysis, which they define as "the modeling of choice from a set of mutually exclusive and collectively exhaustive alternatives." Using transportation demand forecasting as their case study, the authors detail the theories of individual choice behavior, aggregate forecasting techniques, the nested-logit model, and multinomial choice models (among others) with an emphasis on their application to real-world policy planning.

  Past studies conducted by the project team have utilized the discrete choice analysis techniques described in this seminal textbook to estimate the coefficients of its logit market share model. Though this study does not involve a discrete choice analysis component, the project team adapted some of the techniques described in this work are adapted to estimate the coefficients of each customer attribute within its logit model.

- **L. O'Keeffe. "A Choice Experiment Survey Analysis of Public Preferences for Renewable Energy in the United States," Journal of Environmental and Resource Economics at Colby, vol. 01, 2014. [4]**

  In this paper, the author utilizes choice experiment surveys to analyze public preferences for the features of various renewable energy projects. Conditional and mixed multinomial logit models were used to determine the estimates for the important attributes of a project, such as price. The results of the study indicate that while consumers are sensitive to increases in the price of electricity, they also are willing to pay more for projects that are associated with reducing environmental costs.

1.2.2.2.3 Diffusion of Innovations and the Bass Diffusion Model

This section provides the more relevant and seminal papers related to diffusion modeling, with a focus on the Bass diffusion model, which is the method used in this study.

- **F.M. Bass. "A New Product Growth for Model Consumer Durables." Management Science, vol. 15, pp. 215-227, Jan. 1969. [5]**

  Dr. Bass describes the development of his model, which is used to estimate the sales of a product over time. The model assumes that the probability of adoption at any point in time is related to the number of previous adopters. More specifically, Dr. Bass describes the population of potential adopters as either innovators or imitators: innovators being those who will be the first to adopt, and imitators those who adopt based on the signal that those around them are adopting the product.

  The Bass diffusion construct has been applied in scores of studies since its publication in 1969 and was reprinted in 2004 after being identified by *Management Science* as among the Top 10 Most Influential Papers published in the 50-year history of *Management Science*. [6]

- **J.D. Sterman. "The Bass Diffusion Model," in Business Dynamics: Systems Thinking and Modeling for a Complex World, S. Isenberg, New York: McGraw-Hill, 2000, pp. 332. [7]**

In this chapter of the world's leading textbook on system dynamics, Dr. Sterman, the Director of the System Dynamics Group at the Massachusetts Institute of Technology (MIT), takes the Bass diffusion model and describes a few key concepts that improve upon the original model. One such concept is market growth—Bass assumed the market size remained constant, while Sterman describes incorporating market growth. In addition, Sterman details that the entire population will not be interested in purchasing the product given their sensitivity to price, and thus a fraction willing to adopt multiplier is required to adjust the population to a more accurate market size. Finally, Sterman describes integrating a learning curve that informs how the price of a product changes over time as associated with the number of products produced.

The approach utilized in this demonstration study applies the Bass diffusion model using the system dynamics construct described by Sterman. A key advantage of this construct is that it permits any model input to vary over time, including product attributes such as cost, efficiency, or efficacy as well as exogenous factors such as tax credits or incentives, which can change the calculated long-run market share over time. This feature is particularly relevant in the forecasting of solar PV adoption, whose characteristics, especially cost, are rapidly declining as the technology evolves and whose tax environment is also changing over time.

- **V. Mahajan, E. Muller, and Y. Wind. "Diffusion Models, Managerial Applications and Software," in New-Product Diffusion Models. New York: Springer, 2000, pp. 295-310. [8]**

  This book covers several constructs for new product diffusion models, including the Bass diffusion model. A key aspect of this reference is that the authors have estimated the *p* and *q* coefficients of the Bass diffusion model for dozens of technologies using historical product adoption data. While one must be cautious in extrapolating the data from this analysis, it does give practitioners a reasonable range of estimates as a starting point in any analysis, and the values can be used to bound nonlinear optimization routines used to estimate diffusion parameters in other studies, such as this one.

- **B. Sigrin, M. Gleason, R. Preus, I. Baring-Gould, and R. Margolis. "The Distributed Generation Market Demand Model (dGen): Documentation." Internet: https://www.nrel.gov/docs/fy16osti/65231.pdf. Feb. 2016. [9]**

  The National Renewable Energy Laboratory's (NREL's) Distributed Generation Market Demand model provides an example of how Bass diffusion can be utilized to model technology adoption and diffusion. Specifically, the dSolar module simulates the adoption of solar over time as driven by the economic attractiveness of adopting.

- **A. Agarwal. "A Model for Residential Adoption of Photovoltaic Systems." M.S. thesis, California Institute of Technology, California, 2015. [10]**

  Mr. Agarwal's thesis details the application of Frank Bass' diffusion model to the case of solar PV adoption. The model described in the thesis has the flexibility to allow users to input various rate structures, subsidies, and customer demographics to conform to the service region of interest. In addition, this model was trained on Southern California Edison's residential customer data, which included details such as size and date of PV installations, socioeconomic background of the customer, location, and monthly energy consumption.

1.2.2.2.4 Machine Learning Models

This section provides foundational references to support the project team's specification of machine learning (ML) models in this project. Researchers in statistics, computer science, and other fields have published a significant number of works on ML, especially in the past 25 years. The team selected the three references in this section to offer context for the classification and regression tree (CART[1]) and random forest[2] models that were applied in this EPIC project. The third reference, "Classification and Regression by Random Forest," is the reference that most directly relates to the methodology used in this project. The first reference discusses the theoretical foundation and algorithms from which the project methodology can be traced. The second reference compares several of the algorithms spawned from the development in the first reference and that are used in practice today, including the random forest model used for this project.

- **L. Breiman. Classification and Regression Trees. California: Wadsworth & Brooks/Cole Advanced Books & Software, 1984. [11]**

  The methodology used to construct tree structured rules is the focus of this monograph. Unlike many other statistical procedures, which moved from pencil and paper to calculators, this text's use of trees was unthinkable before computers. Both the practical and theoretical sides have been developed in the authors' study of tree methods. CARTs reflect these two sides, covering the use of trees as a data analysis method and in a more mathematical framework, proving some of their fundamental properties. [12]

- **W.L. Loh. "Fifty Years of Classification and Regression Trees." International Statistical Review, vol. 84, pp. 329-348, 2014. [13]**

  The first publication of a regression tree algorithm was in 1963. This paper highlights many of the popular variants on the original algorithm, which are accessible to researchers today through statistical computing platforms like R. Brief summaries of several of the most used algorithms are provided. The paper provides example modeling applications using each of the algorithms presented, with a comparison of the results. Conclusions are provided that offer guidance on pitfalls and algorithm recommendations under different scenarios.

- **A Liaw and M. Wiener. "Classification and Regression by Random Forest." R News, vol. 2/3, pp. 18-22, Dec. 2002. [14]**

  This paper provides an application-focused overview of random forest methodology and its implementation accessible in R through the random forest package. The paper is concise, at just over four pages, and provides short summaries of the algorithm, its usage in the R package, and recommendations for practical usage. Examples are provided for regression, classification, and unsupervised learning.

[1] A classification and regression tree (CART) is a ML algorithm for estimating a numerical or categorical outcome using explanatory variables, with no specification of the assumed structure of the relationship between the outcome and explanatory variables.
[2] Random forest is a ML algorithm that uses reconciliation from estimates from a large number of CART models on sub-samples of historical outcomes and explanatory variables to improve model fit.

The project team's ensemble modeling approach combines CART and random forests where each tree in the forest is an instance of a CART. These trees are trained from historical adoption data. They discover correlations between adoption and a pool of input features derived from a fusion of company and project team data such as measurements derived from utility billing date (e.g., average peak load, baseline load, total daily load, etc.), and different consumer attributes packaged in the project team solution (e.g., demographics, psychographics, premise, and financials). These trees produce adoption profiles that are subsets of key attributes and values that drive varying levels of adoption propensity at the individual customer level.

1.2.2.2.5 Combined Models

The combination of causal methods and ML is a relatively new field of study. The approach employed by the project team in this demonstration project expands upon the approaches previously employed and documented in literature, and appears to be novel in its application of integrating ML with a Bass diffusion and logit modeling construct in forecasting solar PV adoption. This section describes key papers where both causal modeling and ML methods have been applied together.

- **S. Athey and I. Guido. "Machine Learning Methods for Causal Effects." Internet: www.nasonline.org/programs/sackler-colloquia/documents/athey.pdf. 2015. [15]**

  This presentation details the development of a combined causal and ML model as part of a paper that is still a work in progress. The authors present their foundation for developing a new means of estimation that combines a causal approach with machine learning for the prediction component of models, with an emphasis on distinguishing between the causal and predictive parts of the model.

- **P. Bajari, D. Nekipelov, S. Ryan, and M. Yang. "Demand Estimation with Machine Learning and Model Combination." National Bureau of Economic Research, 2015. [16]**

  To develop a tool that would assist econometricians in estimating demand based on large observational datasets, the authors compare statistical analysis methods commonly used to model consumer behavior with the causal methods used in econometric models. They propose a model combining the ML and econometric methods via a weighted linear regression, which was shown to improve the accuracy of a sample prediction of sales data pertaining to salty snacks.

A key element of the modeling approach is the rigor with which the project team assesses the economics of solar PV. The approach selected employs a discounted cash flow optimization model that can evaluate the economics of a solar PV system from the perspective of both the customer and a third-party owner (TPO).[3] [17] A key factor in the rapid uptake of solar PV in the recent past has been the removal of the upfront cost purchase barrier through TPO in combination with a lease, or power purchase agreement (PPA), contract structure. Though the pendulum of system ownership appears to be swinging back toward

[3] "Third-party financing is a well-established financing solution in the United States, having emerged in the solar industry as one of the most popular methods of solar financing. Third-party solar financing predominantly occurs in two forms: solar leases and power purchase agreements (PPAs) In the lease model, a customer signs a contract with an installer/developer and pays for the use of a solar system over a specified period of time, rather than paying for the power generated. In the PPA model, the solar energy system offsets the customer's electric utility bill, and the developer sells the power generated to the customer at a fixed rate, typically lower than the local utility." www.epa.gov/repowertoolbox/understanding-third-party-ownership-financing-structures-renewable-energy.

customer ownership, TPO is expected to continue to play a large role in the adoption of distributed solar PV and thus, must be accounted for in a rigorous adoption model; however, any ownership construct (cash purchase, financed, TPO) could be calculated in this model. Discounted cash flow techniques are not new, though their application to forecasting solar PV and their ability to help explain pricing strategies of TPO system providers is particularly relevant in the solar PV technology space.

- **F. Stermole and J. Stermole. Economic Evaluation and Investment Decision Methods. Investment Evaluation Corporation, 2012. [18]**

  This textbook provides a comprehensive overview of how to perform economic evaluation under many different scenarios. Using examples with respect to industries such as oil & gas, mining, and energy, this textbook provides the fundamentals of computing discounted cash flows.

- **Navigant Consulting, Inc. *Solar Project Return Analysis for Third Party Owned Solar Systems.* 2016. [19]**

  As part of the analysis performed to evaluate the TPO solar PV leasing business model in Arizona, Navigant utilized its RE-SIM™ model to conduct discounted cash flow analysis and to compare the economics of TPO systems in different service territories. The cash flow streams accounted for in this analysis included initial capital outlay, debt-financing cash inflow and interest payments, incentives, accelerated depreciation for tax purposes, Federal Investment Tax Credit benefits, and much more. The RE-SIM model is a nonlinear optimization model with the objective to minimize the lease or PPA rate within the bounds of the constraints. The analysis conducted in this study helped to shed light on pricing strategies of TPO PV providers, which is critical to understanding the impact of forecast cost declines or tax credit changes and to forecast product adoption. This study was submitted as part of formal written and oral testimony in the 2016 UniSource Electric rate case (Docket Number E-04204A-15-0142).

- **U. Benzion and J. Yagil. "Decisions in Financial Economics: An Experimental Study of Discount Rates," Advances in Financial Economics, M. Hirschey, J. Kose, A.K. Makhija, Eds. United Kingdom: Emerald Group Publishing, 2002, pp. 19-40. [20]**

  The authors perform an experimental study involving individuals of varying economic understanding, finding that implicit discount rates decrease as the time horizon or monetary sum decrease. In addition, they find that implicit discount rates approach market interest rates as the economic understanding or formal education of the subjects increases.

  Their results also indicate that across all combinations of the subject's education, monetary sum, and time horizon, people tend to have inherently high implied discount rates. These high implied discount rates are captured in the approach used for this demonstration project, which allows for sensitivity analyses to be performed around variables such as customer discount rate, incentive levels, and payback period as well as the resulting adoption of solar PV.

1.2.2.3 Outcome

After careful review of the literature cited above, the project team identified the methodological framework to be demonstrated in this project using combined methods. The project team chose to use a Bass diffusion system dynamics causal model (BDCM) to demonstrate the methodological framework.[4] While multiple options exist for forecasting adoption of any product over time, this approach combines several well-established techniques, including Bass diffusion, random utility theory, and system dynamics causal modeling, and provides a robustness and flexibility not present in more simplistic approaches (e.g., regression of a logistic curve). When combined with more recently developed ML algorithms, which can identify those attributes that are most salient, the methodology offers even greater potential to provide comparatively accurate granular adoption forecasts (e.g., at the ZIP code level or possibly lower if more granular data is provided).

1.2.3 Subtask 3.2: Methodology Framework Development

1.2.3.1 Objective

Leveraging the literature review, develop a methodological framework to model customer propensity to adopt PV systems.

1.2.3.2 Approach

This section provides a high-level overview of the methodology selection process for this demonstration study. The BDCM model uses a discrete choice market share approach in conjunction with a calibrated Bass diffusion model to forecast the adoption of PV or other DER technologies. The approach considers situations in which a consumer selects one option from a finite set of alternatives (e.g., the TPO solar PV business model presents a choice between two rates of electricity to consumers). In the discrete choice portion of the framework, the decision maker chooses the solution that maximizes a utility function that depends on several economic and non-economic attributes.

The BDCM model components simulate the S-shaped approach to equilibrium that is commonly seen for technology adoption. In this classic application of the Bass model, market potential adopters flow to adopters through two primary mechanisms: adoption from external influences such as marketing and advertising, and adoption from internal influences or word of mouth. Table 1 provides a high level summary of the key inputs and outputs of BDCM. Table 1 provides a stock/flow diagram illustrating the causal influences underlying the BDCM model, along with typical examples of adoption S-curves.

[4] Navigant Consulting, Inc.'s proprietary RE-SIM™ model that uses a discrete choice market share approach in conjunction with a calibrated Bass diffusion model to forecast the adoption of PV or other DER technologies.

Table 1. BDCM Key Inputs and Outputs

Key Inputs	Key Outputs
• Technology cost and performance forecast • Electricity prices and utility offset rates • Utility incentives, state and federal tax credits • Historical installed capacity, building stocks, consumption • Market diffusion parameters and consumer sensitivity • Eligibility constraints (homeownership rates, PV access factors, etc.)	• Levelized cost of energy • Levelized value of electricity • Price response curves • System-level technical and market potential forecast • Potential as a percentage of sales • Calibrated (through back-casting) market diffusion parameters

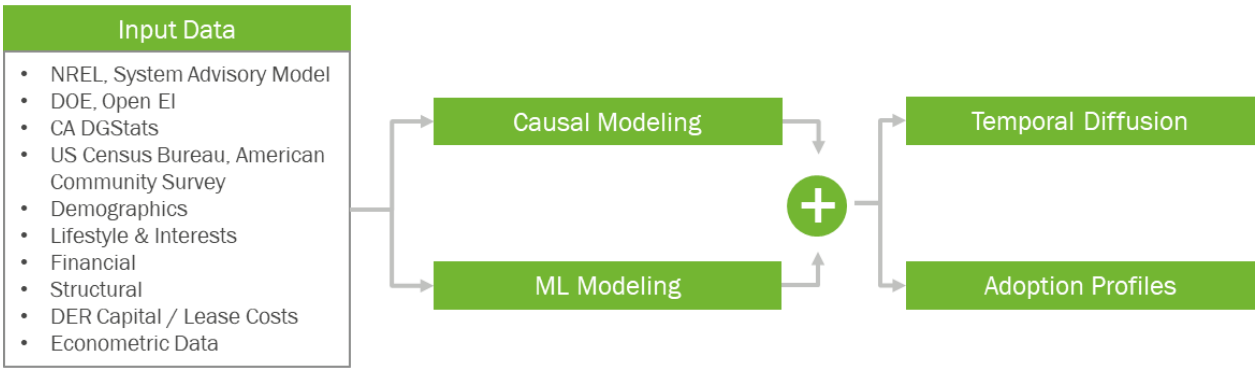
1.2.3.2.1 Adoption Forecasting Model

The adoption forecast model combines the BDCM model and the ML method to create additional value for customer analytics. This joint modeling approach augments the well-established discrete choice causal models that are largely driven by subject matter expertise, with additional adoption information that is automatically discovered through a data-driven and ML process. The net effect of this approach, depicted in Figure 2, provides a more accurate market potential estimate, along with additional insights gained from customer-specific adoption propensities. The synthesis of ML and BDCM occurs through fine-tuning the long-run market share calculations in BDCM with ZIP code-level customer characteristics from ML. More specifically, the non-economic attributes are further delineated by the important attributes identified by ML.

1.2.3.3 Outcome

The results of this framework identification solidified the approach used for the demonstration and identified key inputs and outputs of the model.

Figure 2. Adoption Forecasting Methodology for SDG&E



1.2.4 Subtask 3.3: Demonstration

1.2.4.1 Objective

The goal of this task was to demonstrate the methodology framework developed using San Diego Gas & Electric (SDG&E) data.

1.2.4.2 Approach

1.2.4.2.1 Demonstration Use Case

As a use case, the project team forecasted solar PV adoption at a ZIP code level, for all ZIP codes specific to SDG&E (118 ZIP codes). Specifically, the project team utilized the California Distributed Generation Statistics Currently Interconnected Data Set [21], with *App Received Date* serving as the proxy for the installation date of the net energy metered (NEM) systems. The project team modeled solar PV adoption for the residential sector in each of the applicable ZIP codes over a 20-year time horizon (1998-2017). The ML analysis identified the key customer attributes that factor most prominently into PV adoption, accounting for all NEM installations across the investor-owned utility (IOU) territories in California. Using the larger dataset for the ML component of the analysis, as opposed to only the SDG&E ZIP codes, facilitated a better understanding of which customer attributes factor most prominently into adoption of solar PV. This is because it provides additional variation in the candidate model covariates for the ML models to train on from the more than 1,000 additional ZIP codes in the Southern California Edison (SCE) and Pacific Gas & Electric (PG&E) service territories.

1.2.4.2.2 Data Sources

The demonstration use case does not rely on any granular or customer-level data from SDG&E. Rather, the project team utilized publicly available data in concert with data from other studies conducted by the project team in other jurisdictions, which, in some cases, may be proprietary. Any proprietary data utilized demonstrated the methodology and output but was not provided in raw form to SDG&E as part of any deliverable. Table 2 provides a list of the public data used for this demonstration.

Table 2. Public Data Sources Used for this Demonstration

Data	Source	Usage in Model
Solar PV Net Metering Interconnection	California Distributed Generation Statistics, http://www.californiadgstats.ca.gov/downloads/	To calibrate diffusion and market share parameters and to compare historical adoption[5] with simulated adoption
Residential Load Profiles	US Department of Energy, Open EI, https://openei.org/doe-opendata/dataset/commercial-and-residential-hourly-load-profiles-for-all-tmy3-locations-in-the-united-states	Not used in the model—used to calculate residential offset rate in SDG&E territory
Solar PV System Production	NREL, System Advisory Model, https://sam.nrel.gov/	Not used in the model—used to calculate residential offset rate in SDG&E territory
American Community Survey	US Census Bureau, https://factfinder.census.gov/	To compile ZIP code-level demographic attributes, which served as candidate covariates in the ML models
2010 Decennial Census	US Census Bureau, https://factfinder.census.gov/	To estimate households by ZIP code
California Energy Commission	Building Climate Zone by ZIP Code, http://www.energy.ca.gov/maps/renewable/BuildingClimateZonesByZIPCode.pdf	To map climate zones to ZIP code
State of California Dept. of Finance	P1: State Population Projections (2010-2060), http://www.dof.ca.gov/Forecasting/Demographics/projections/	To forecast 2010 Census Data through the 2017 calibration period

1.2.4.2.3 Parameter Development

The project team determined which model parameters/coefficients were assumed to be fixed versus which parameters were solved for using available data to calibrate the model (e.g., using nonlinear optimization techniques). The project team estimated key model parameters by leveraging information from other analyses, publicly available data, and proprietary datasets regarding customer attributes in each ZIP code. The project team approach for estimating key model parameters is detailed in the following sections.

Key Customer Attributes

In this project, the primary purpose of ML was to discover the most important attributes that drive PV adoption at the ZIP code aggregate level. The project team tuned the model settings to distill the attributes that were most important for defining PV adoption propensity, along with the structural relationship between those attributes. ML models in the CART family are well-suited for this project because it is not well known which consumer attributes are associated most with PV adoption. ML

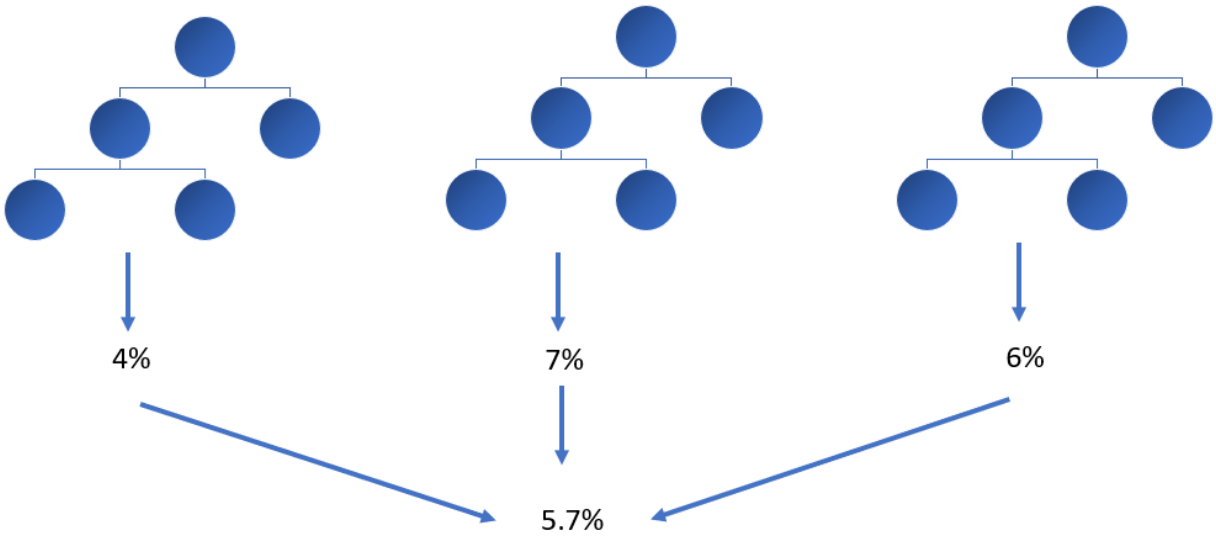
[5] The team used the *Application Received* date for the time field, though future analysis should use the *Application Complete* date, which is time lagged by about a month relative to the *Application Received* date.

algorithms are an efficient means for discovering those attributes, especially when missing values are not prevalent in the analysis of the ZIP code-level data. Random forest ensemble modeling offers robust identification of the most important attribute drivers compared with single models that risk overfitting to the data. The suitably large number of ZIP codes in California IOU territories makes the subsampling aspect of the random forest algorithm appropriate.

The project team used random forest ML models in this demonstration, where each tree in the forest is an instance of a regression tree for predicting the proportion of customers in a ZIP code that have net meters installed for PV systems, a continuous variable. These trees are trained from historical adoption data, discovering correlations between adoption and a pool of input features derived from a fusion of Census Bureau and project team customer attribute data (e.g., demographics, psychographics, premise physical characteristics, and financials).

Figure 3 illustrates the prediction from multiple trees in an ensemble being aggregated and reconciled into a single estimate.

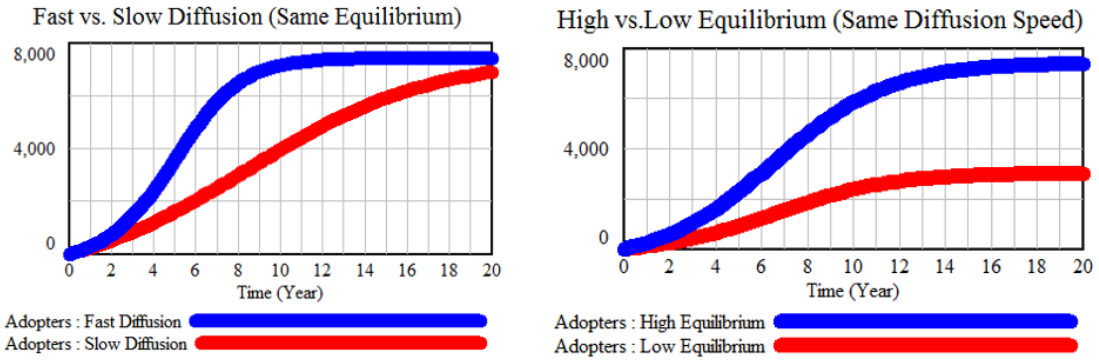
Figure 3. Ensemble Model: Example for Regression



Adoption Parameters

Estimating PV adoption using the enhanced Bass diffusion model can be broken into two parts: calculating long-run market share and the rate at which the long-run market share is achieved over time. The left graph in Figure 4 illustrates two adoption profiles approaching the same long-run market share but at different rates of diffusion, which are governed by the *p* and *q* parameters of the Bass diffusion model (there are also referred to as marketing effectiveness and work of mouth strength in the BDCM adoption model). In contrast, the right graph in Figure 4 illustrates two adoption profiles with different long-run market share but constant rates of diffusion (i.e., constant Bass diffusion coefficients).

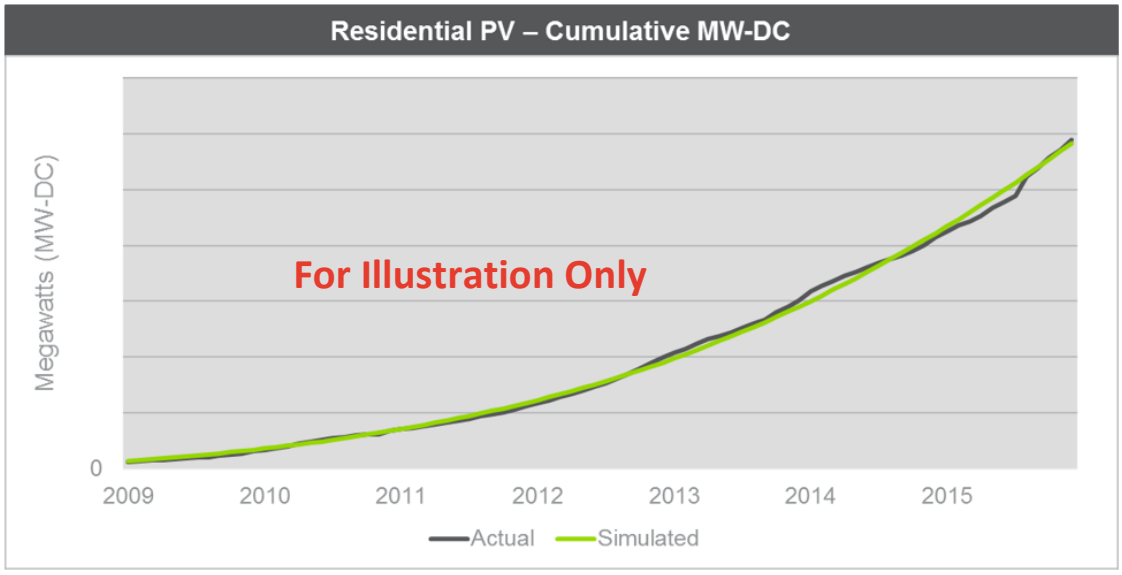
Figure 4. Illustrating Difference Between Rate of Diffusion and Long-Run Market Share



Unless the population is sufficiently far along in the adoption S-curve (e.g., past the inflection point of adoption), it is typically impossible to simultaneously estimate both the long-run market share and the Bass diffusion coefficients using historical data alone. This is because an infinite combination of coefficients tends to yield equally good data fits. As such, one must typically fix one set of parameters while calibrating the coefficients of the other.

The literature review revealed multiple cases that utilized a fixed functional format for calculating long-run market share (informed through a combination of means, including but not limited to discrete choice analysis), typically using a logit market share model. The project team calibrated the Bass diffusion coefficients of marketing effectiveness and word of mouth strength (the *p* and *q* in the Bass model) through nonlinear optimization techniques to achieve the best possible fit for simulated and historical adoption. Figure 5 is for demonstration purposes only and illustrates a diffusion coefficient calibration, showing a close fit between historical adoption and simulated adoption in the residential sector of a utility service territory. This calibration process entailed adjusting Bass diffusion coefficients through nonlinear optimization to minimize the absolute value of the difference between simulated and actual adoption totaled over each month of the simulation forecast.

Figure 5. Illustration of a Calibration of Bass Diffusion Coefficients – Comparison of Actual Historical vs. Model Simulated Adoption



In this demonstration, the project team was primarily interested in fine-tuning the long-run market share calculations to facilitate obtaining a more granular forecast of spatial adoption, in this case by ZIP code. The project team sought to maximally leverage information obtained through similar studies conducted by the project team in California and Arizona and to calibrate coefficients of the logit long-run market share model and the diffusion model. The approach seeks to marry the granular data and ML methods with causal Bass diffusion modeling without having to resort to more costly analysis approaches such as executing customer discrete choice analysis surveys. The survey approach could be used in future analyses to facilitate independent estimation of both the long-run market share parameters and the Bass diffusion coefficients; however, it was outside the scope of this demonstration.

Logit Long-Run Market Share Coefficients

The functional format of the binary logit model used to estimate long-run adoption of solar PV in a utility service territory is illustrated in Equation 1.

Equation 1. Binary Logit Market Share Model

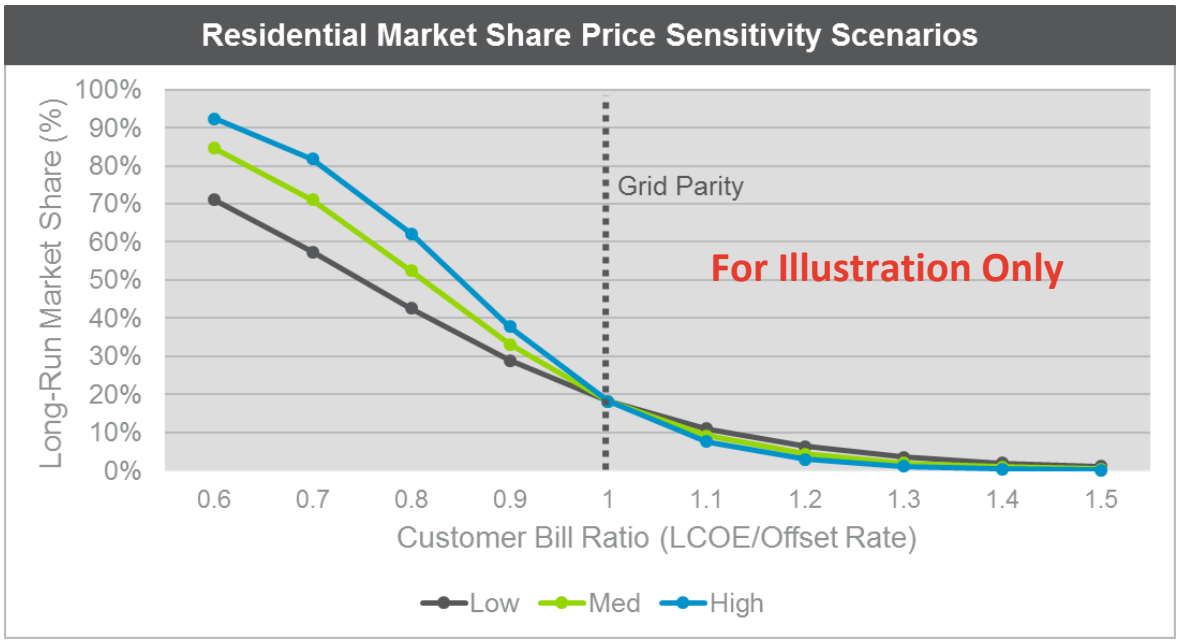
$$Long\ Run\ Solar\ PV\ Market\ Share = \frac{1}{1 + e^{(\alpha + \beta \cdot Price\ Ratio)}}$$

Where, the price ratio was calculated to be the ratio between the lease rate, or PPA rate—typically calculated as a levelized cost of electricity (LCOE)—that could be offered by a TPO of a PV system and the utility's electricity offset rate for a customer with a PV system installed.[6]

[6] Utility offset rates ($/kWh) are defined as the dollar value of a customer's bill reduction for each kilowatt-hour generated by the customer's solar system. It is the amount of their bill that is offset for each kilowatt-hour generated (hence the term). In other words, it is the amount a customer saves on their utility bill.

The shape of the above logit market share model was like the illustrative graphic, demonstrating three separate β coefficients.

Figure 6. Logit Market Share Model



In this demonstration, the project team utilized a logit market share model by expanding the α coefficient, sometimes referred to as an alternative specific coefficient, to include the top four to six customer attributes identified by the project team's ML analysis.

Thus, the project team further disaggregated the α coefficient (sometimes referred to as an intercept term) as follows in Equation 2, holding the price coefficient constant.

Equation 2. Disaggregation of the Intercept Term

$$\propto = \propto_0 + \propto_1 \cdot \boldsymbol{Attribute_1} + \propto_2 \cdot \boldsymbol{Attribute_2} + \propto_3 \cdot \boldsymbol{Attribute_3} + \cdots$$

The project team applied statistical methods to calculate the best fit for each of the disaggregated α coefficients using the combination of the following by ZIP code: historical adoption, number of households, aggregate customer attributes (e.g., median income), and a scaling factor (calculable using the fixed Bass diffusion parameters). The team used these to account for how far along on the adoption S-curve each ZIP code should theoretically be as a percentage of maximum adoption.

1.2.4.3 Outcome

The outcome of the demonstration is detailed in Section 2.

1.2.5 Subtask 3.4: Disadvantaged Communities Analysis

1.2.5.1 Objective

The goal of the DAC demonstration was to conduct ML analytics on DAC ZIP codes, model DAC ZIP code solar propensity to adopt and compare the results to non-DAC ZIP codes in California.

1.2.5.2 Approach

1.2.5.2.1 Disadvantaged Community Definition

Disadvantaged Communities (DAC) are communities designated by CalEPA, pursuant to Senate Bill 535, using the California Communities Environmental Health Screening Tool (CalEnviroScreen). DAC are identified by census tract and include the tracts that scored at or above 75% in the 3.0 version of the CalEnviroScreen that is available at the time of the EPIC application. [22] For the purposes of this analysis, the project team defined a DAC ZIP code as a ZIP code where >50% of the population lives in census tracts with a CalEnviroScreen 3.0 score > 75%. Using the SB 535 Disadvantaged Communities List, 255 ZIP codes in California meet this criterion; however, only two of these ZIP codes were in SDG&E territory, requiring this analysis to be conducted at the state level.

1.2.5.2.2 Data Sources

The DAC demonstration use case used the same data that was used in the prior subtask, with the additions provided in Table 3.

Table 3 Additional Public Data Sources Used for DAC Demonstration

Data	Source	Usage in Model
List of Disadvantaged Communities	http://www.energy.ca.gov/commission/diversity/definition.html	Used to determine DAC and non-DAC ZIP codes for ML, statistical and propensity to adopt analysis.

1.2.5.2.3 Parameter Development

For the DAC analysis, the project team determined which model parameters/coefficients were assumed to be fixed versus which parameters were derived using available data to calibrate the model (e.g., using nonlinear optimization techniques). Key model parameters were estimated by leveraging information from other analyses, publicly available data, and proprietary datasets regarding customer attributes in each ZIP code. The approach for estimating key model parameters is detailed in the following sections.

Key Customer Attributes

In the DAC analysis, the primary purpose of ML was to discover the most important attributes that drive solar PV adoption at the ZIP code aggregate level, specifically within DAC-designated ZIP codes. Since there were only two ZIP codes in SDG&E territory with a DAC population >50%, the project team applied its ML models to the 171 DAC ZIP codes in the IOU service territories with Census Bureau demographics data were available. As with the primary analysis unrestricted to DAC ZIP codes, CART and random forest ensemble models were used to distill the attributes most strongly linked to PV adoption percentage at the ZIP code level.

Statistical Analysis

The project team ran statistical analyses on the key customer attributes to gain insight into the differences in these attributes in DAC versus non-DAC ZIP codes. Specifically, the team tested the mean difference of the attribute value for statistical significance, and compared the distribution of the attributes across the DAC and non-DAC ZIP codes. In addition, the team also analyzed the percentage of owner occupied homes, and solar PV market share.

Adoption Parameters

The project team employed the same diffusion modeling methodology for the DAC analysis as described in Sections 1.2.3 and 1.2.4. While the construct was the same, the region analyzed differed in that it included all California ZIP codes for which data were available, as opposed to limiting the analysis to the SDG&E service territory. The primary reason for expanding the analysis region was to provide an adequate sample size for both DAC and Non-DAC ZIP Codes (SDG&E's service territory only has twos ZIP codes that met the DAC definition outlined in Section 1.2.5.2.1.). Additionally, the optimization of diffusion parameters aggregated applicable ZIP code adoption data to provide two sets of parameter estimates – one for the aggregation of all DAC ZIP codes and one for the aggregation of all non-DAC ZIP codes.

1.2.5.3 Outcome

The outcome of the DAC demonstration is detailed in Section 2.

1.2.6 Task 4 – Conduct Data Analysis and Develop Findings and Recommendations

1.2.6.1 Objective:

Conduct data analysis and develop findings and recommendations.

1.2.6.2 Approach

The project team performed data analysis and developed results and recommendations on next steps as outlined in Section 4 of this report.

1.2.6.3 Output

Data analysis, findings and recommendations, as provided in Section 3 and Section 4 of this report.

1.2.7 Task 5 – Final Report

1.2.7.1 Objective

Aggregate all findings and compile it into a comprehensive report.

1.2.7.2 Approach

Develop a comprehensive final report based on an agreed-upon outline developed by the team. The project team used results from the demonstration and data analysis to develop the final report. The report was prepared as a draft for review by project stakeholders and then revised into final form, based on the review comments.

1.2.7.3 Output

Comprehensive final report as presented in this document.

2. Results and Analysis

2.1 Demonstration Activity Results and Analysis

The following sections detail the results of each stage of the demonstration activity. As noted in Section 1.2.3, the demonstration plan included the analysis of PV adoption in all ZIP codes across the three California IOU service territories to provide approximately 10 times the ZIP codes of data to train the models that would then be applied to the 118 ZIP codes in the SDG&E service territory.

2.1.1 Customer Attributes Driving Adoption

The project team conducted ML modeling to identify the most important attributes driving adoption at the ZIP code level. This analysis included compiling the analysis data and fitting it to a random forest ML model.

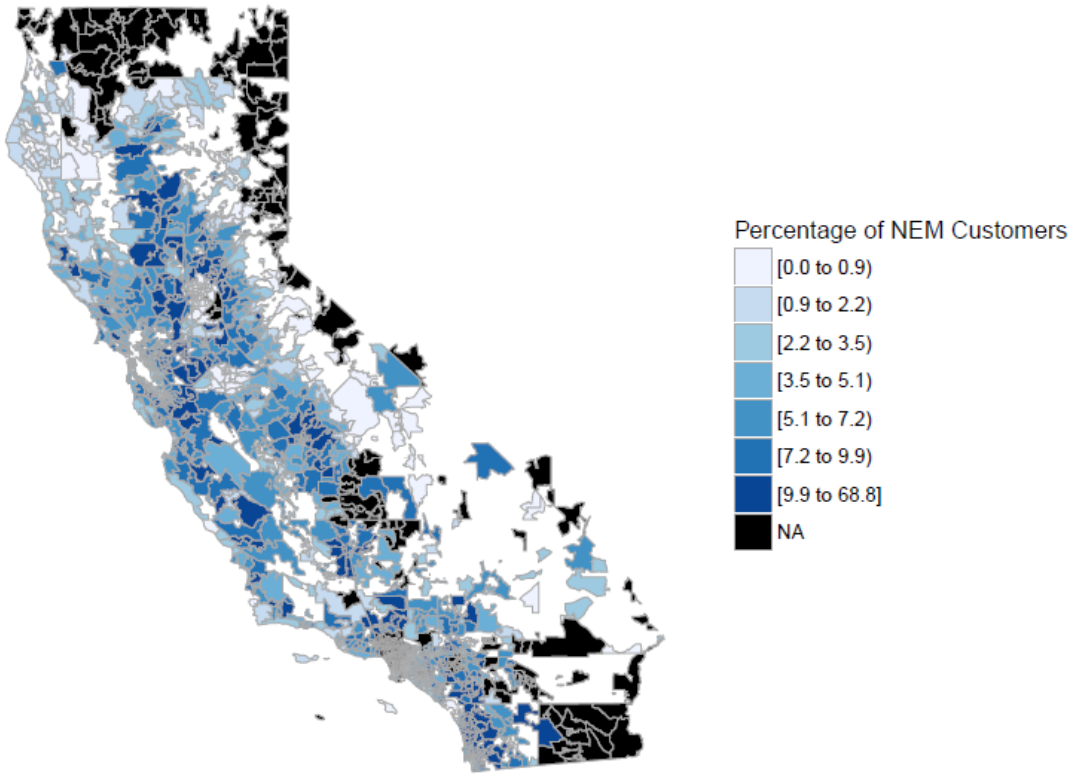
2.1.2 Data Compilation

The project team imported solar PV interconnection data using the California Distributed Generation Statistics (DGStats) NEM interconnections database. [23] This dataset contains a record for each individual NEM solar PV install for customers in all IOU territories (e.g., SDG&E, SCE, and PG&E), using the ZIP code as the most granular customer identifier. The project team filtered the dataset to include only residential customers. The team then aggregated the data to the ZIP code level by month and year for the granular adoption analysis and cumulatively over all the years for the customer attribute analysis.

The project team leveraged demographics and physical attributes available from the US Census Bureau at the ZIP code level, incorporating climate zone identifiers at the ZIP code level from the California Energy Commission to account for weather differences. In total, the team assembled approximately 100 covariates for use in the ML modeling process.

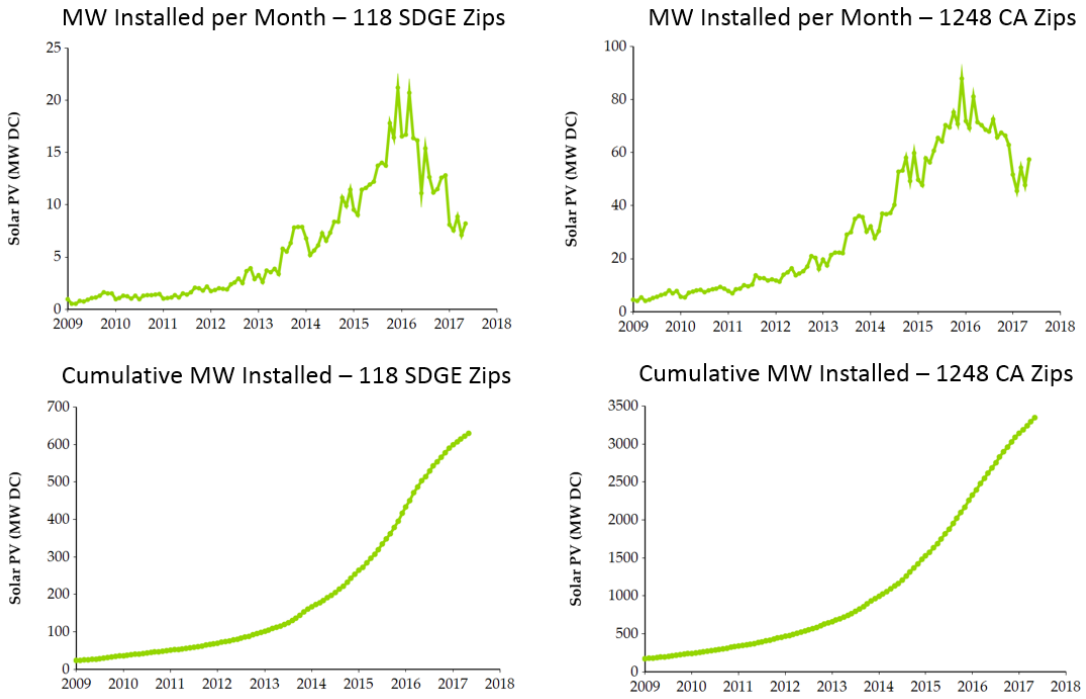
The team converted total NEM installs by ZIP code into the proportion of households to establish a common analysis basis across the ZIP codes, which vary by size. This necessitated identifying ZIP codes that make up the three IOUs in the state to account for ZIP codes without any NEM installs in the DGStats database. The final data preparation step involved filtering the data to include customer demographic data in only those ZIP codes where data is readily available across all fields. The resulting dataset covered 1,248 out of 1,659 ZIP codes represented in the DGStats database and 116 of 118 ZIP codes specific to SDG&E's service territory, as presented in Figure 7. These data represented roughly 12 million homes in California and 1.4 million homes in SDG&E's service territory. The two ZIP codes in the SDG&E territory without an estimate did not have any residential households, according to Census data.

Figure 7. NEM Installation Percentage by ZIP Code



Next, the project team aggregated system installation data by month of install to enable modeling over time rather than for the current snapshot through 2017, as was done for the ML modeling. Figure 8 shows the aggregate adoption data for the 1,248 ZIP codes modeled in California and the 118 ZIP codes modeled in SDG&E's service territory, both on an incremental (i.e., new installations per month) and cumulative megawatt installed basis, including data through the first quarter of 2017.

Figure 8. Incremental and Cumulative MW Installed in Modeled ZIP Codes – Demo Results[7]



2.1.3 Random Forest Model

The project team fit a random forest model consisting of individual CART models applied to each of 500 random subsets of the analysis dataset of California ZIP code NEM installation data and corresponding covariates. The project team chose to use 500 random subsets, which was determined to be sufficient because additional subsamples showed diminishing value to the explanatory power of the model. The resulting model accounted for 48% of the variation in the ZIP code-level proportion of NEM installations across the IOUs and 76% of the variation for the SDG&E ZIP codes.

Figure 9 depicts the random forest error distribution for estimating the proportion of households with NEMs for PV system installations—i.e., the error of the ML model to the actual DGStats data. The green curve shows the distribution of the estimation error for the SDG&E ZIP codes, and the black curve shows the estimation error distribution for other IOUs in California. For both the SDG&E and other IOU ZIP codes, the error distribution is tightly centered around zero.

[7] Historical installation data aggregated from DGStats.com, filtered to include only residential installations.

Figure 9. Random Forest Estimation Error Distribution for Percentage of Households with NEM Installations

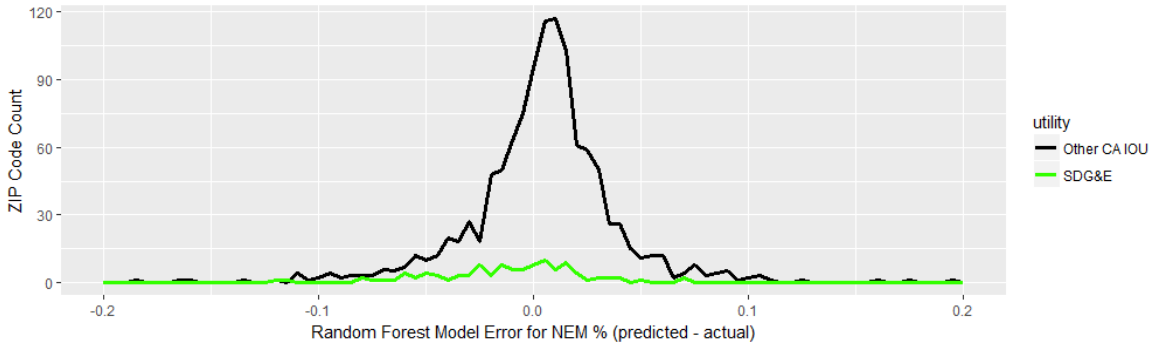


Table 4 details the numerical summary of the error distribution shown in Figure 9. For SDG&E, a slight skew toward underestimating NEM installation proportions occurs, as both the mean and median were less than zero. Of SDG&E ZIP codes, 50% had NEM installation percentage estimates between -2.8% and 0.8% of the true installation percentage.

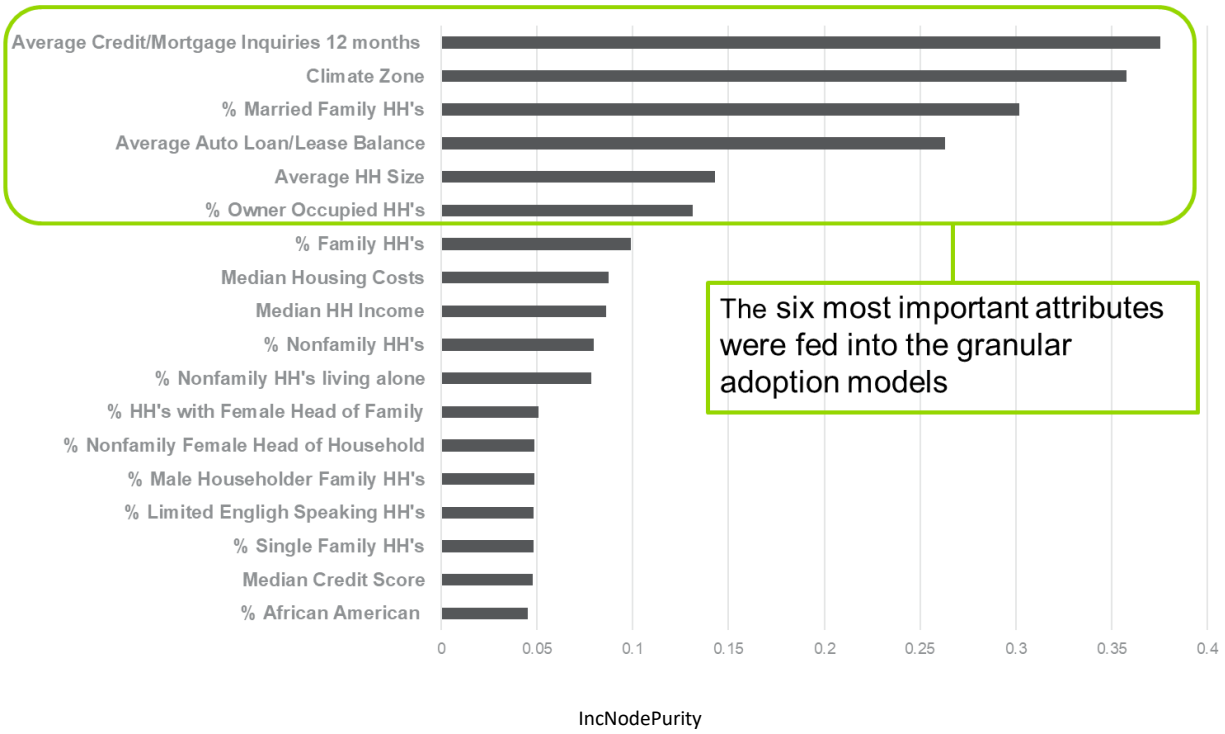
Table 4. Random Forest Estimation Error Quartiles and Mean Percentage of Households with NEM Installs

Utility ZIP Codes	Min	1st Quartile	Median	Mean	3rd Quartile	Max
SDG&E	-12.0%	-2.8%	-0.6%	-1.1%	0.8%	6.8%
Non-SDG&E	-57.6%	-1.0%	0.5%	0.2%	1.9%	19.5%

While the 76% reduction in the NEM percentage variance in SDG&E ZIP codes was an important validation for the random forest model goodness of fit, the primary purpose for the random forest model in this project was to isolate the most significant ZIP code-level drivers for estimating NEM adoption percentage.

The project team examined the variable importance measure from the random forest model, shown in Figure 10. The variable importance in the x-axis of the plot, labeled IncNodePurity, is the relative degradation in model precision associated with the removal of the listed model covariate. Figure 10 shows six attributes, enclosed by the green box, that stick out significantly to the right; as a result, these were determined to be the most important attributes estimating ZIP code-level PV NEM adoption.

Figure 10. Variable Importance Plot – Estimating Percentage of NEM Installation Households



These attributes correlate to varying degrees with the approximately 100 candidate covariates, of which the top 30 in terms of importance are shown in Figure 10. The six attributes in the green box in Figure 10 and listed in Table 5 in order of most to least important were the most predictive for estimating NEM percentage.

Table 5. Key Customer Attributes

Key Customer Attributes
Average number of credit mortgage type inquiries in the last 12 months
Climate zone
Percentage of households that are married families
Average balance on an open auto loan and lease trades reported in the last 6 months
Average household size
Percentage of owner-occupied housing units

2.1.4 Granular Adoption Analysis

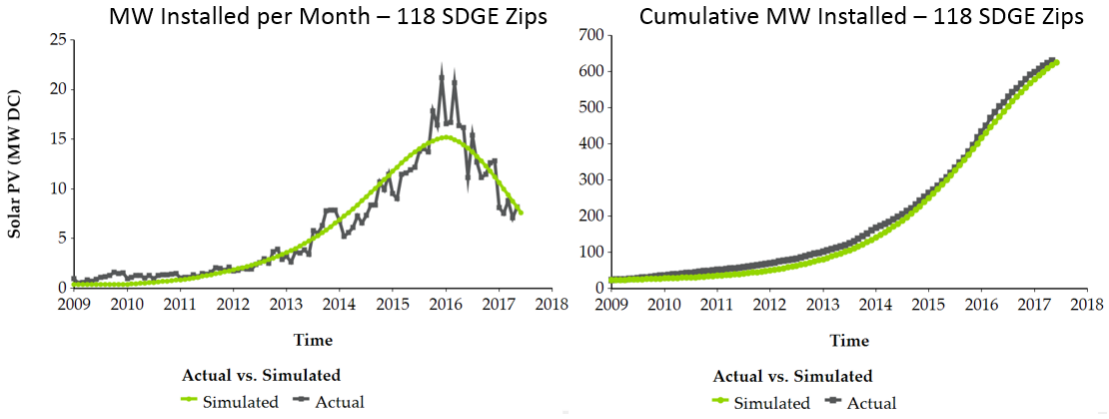
This section presents the results of analysis that explores the ability of the project team's causal model to replicate historical adoption profiles, which provide a greater level of confidence that such a model may provide reasonable forecasts moving forward. This section provides a description of a three-step analysis process, each of which improves the ability of the causal model to replicate historical adoption at a granular level (e.g., ZIP code level).

2.1.4.1 Aggregate Adoption Analysis – Statewide Coefficients

The project team began the adoption analysis by first running a nonlinear optimization to solve for the best fit of its model parameters for adopting data in SDG&E's service territory. The project team estimated the two parameters (marketing effectiveness and word of mouth, or the *p* and *q* in the Bass model) that would drive long-run (equilibrium) adoption in the logit market share model (discussed in Section 1.2.3), assuming a historical price ratio of 0.8 over the historical period simulated. While this is a somewhat simplified assumption, evidence suggests that TPOs tend to price leases and PPAs based on prevailing electricity rates, and that they have been able to historically price below utility electricity rates since introducing the TPO business model. [24] Initially, the plan was to hold the Bass diffusion parameters constant in accordance with values the team has estimated from other studies. However, the adoption data suggests that the market may be past the inflection point of the typical S-curve adoption profile, which means that more information is available to estimate diffusion parameters than originally anticipated. Through a calibration process, the project team independently estimated the two Bass diffusion parameters (*p* and *q* in the classical model, or advertising effectiveness and word of mouth parameters in the team's BDCM model), in addition to the parameters governing the long-run market share. All parameters in this step of the analysis were held constant across all ZIP codes modeled, though the team did back-cast historical adoption for each ZIP code.

The project team obtained an excellent fit of the data when aggregated across all ZIP codes, both on a cumulative megawatt installed basis and on an incremental monthly installation basis. Figure 11 provides a graphical view of the model simulated adoption compared with the actual adoption data gathered from the DGStats database. This result suggests that the Bass diffusion modeling approach is well-suited for analyzing the adoption of solar PV in the residential market, whose dynamics of market growth and saturation appear to be consistent with those observed with numerous other successful product deployments.

Figure 11. Model Simulated vs. Actual PV Installed – Statewide Coefficients Only

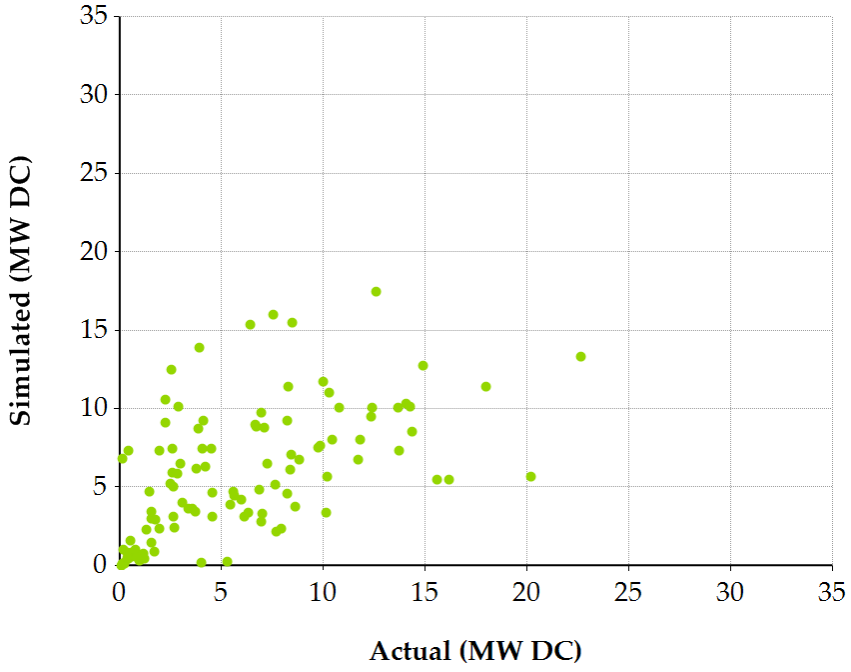


The project team noted that to obtain the strong fit seen above, it added an additional dynamic into the Bass model. Specifically, the team permitted the word of mouth parameter to grow over time, which results in a somewhat steeper (or super-exponential) growth profile. Using only a single, constant value for this parameter resulted in a flatter bell curve of incremental adoption and did not fit the data as well. The implications of this model modification are that the positive feedbacks that tend to generate the exponential growth early in the adoption cycle have been accelerating over time. Such positive feedbacks

include the word of mouth effect but are not limited to that feedback since the model fit will effectively load onto this single parameter any other positive, or reinforcing, feedback that exists in the actual market dynamics. With the growth of social media over the last decade, it is reasonable to believe that this social networking parameter has indeed grown over time. Considering the limited scope and timeframe of the project, the team does not assert that this is the only mechanism by which a super-exponential type of growth dynamic could be caused.

Because forecasting adoption at a more granular level than the entire residential sector was a focus of this project, the model included a back-cast of adoption for each of the 118 ZIP codes the team modeled in SDG&E's service territory. However, without introducing additional model parameters to explain observed variance in adoption across ZIP codes, the project team did not anticipate getting a tight back-cast fit when disaggregating to the ZIP code level. Figure 12 shows the model simulated versus actual cumulative megawatts of adoption from January 2009 through March 2017 for each of the 118 ZIP codes modeled. As seen in this figure, where each data point represents a different ZIP code, substantial noise exists in the adoption back-cast at the ZIP code level. A better data fit would show each data point falling along the diagonal of the graph, if simulated and actual adoption were closely correlated.

Figure 12. Simulated Cumulative Adoption (2009-2017) for all 118 SDG&E ZIP Codes – Statewide Coefficients Only (Held Constant Across All ZIP Codes)



The above results warrant the addition of model parameters to better explain the variance from ZIP code to ZIP code.

2.1.4.2 Granular Adoption Analysis – Incorporating Customer Attributes

The next step of the team's analysis entailed incorporating the customer attributes calculated in the ML analysis described in Section 2.1.1. Observation of ZIP code-level adoption data suggested that long-run market share across ZIP codes varied more substantially than the rate of diffusion that would be governed by the Bass diffusion parameters (advertising effectiveness and word of mouth). As a result, the project

team held these diffusion parameters constant across all ZIP codes and incorporated customer attribute coefficients into its logit market share model, which effects the calculation of long-run market share.

As discussed in Section 1.2.3, the team expanded the two-parameter logit market share calculation into Equation 3

Equation 3. Adding Customer Attribute Coefficients to the Long-Run Market Share

$$LongRun\_MarketShare_{ZipCode} = \frac{1}{1+e^{\alpha_{ZipCode}+\beta \,\mathrm{Price}Ratio}}$$

Where,

$$\alpha_{ZipCode} = a1_{ClimateZone} + a2*attribute2_{ZipCode} + a3*attribute3_{ZipCode} + a4*attribute4_{ZipCode} + a5*attribute5_{ZipCode}$$

The above function effectively resulted in a logit market share calculation having the following number of coefficients:

- Price coefficient (β): One global price coefficient
- Climate zone-specific coefficients (a1): 16 coefficients, one for each climate zone in California
- Customer attribute coefficients (a1-a5): Five coefficients, one for each of the five key customer attributes, the value of which varies by ZIP code:
    - $attribute2_{ZIPCode}$ = Average number of mortgage-type credit inquiries in the last 12 months
    - $attribute3_{ZIPCode}$ = Percentage of households that are married families
    - $attribute4_{ZIPCode}$ = Average balance on open auto loan and lease trades reported in the last 6 months
    - attribute $5_{ZIPCode}$ = Average household size

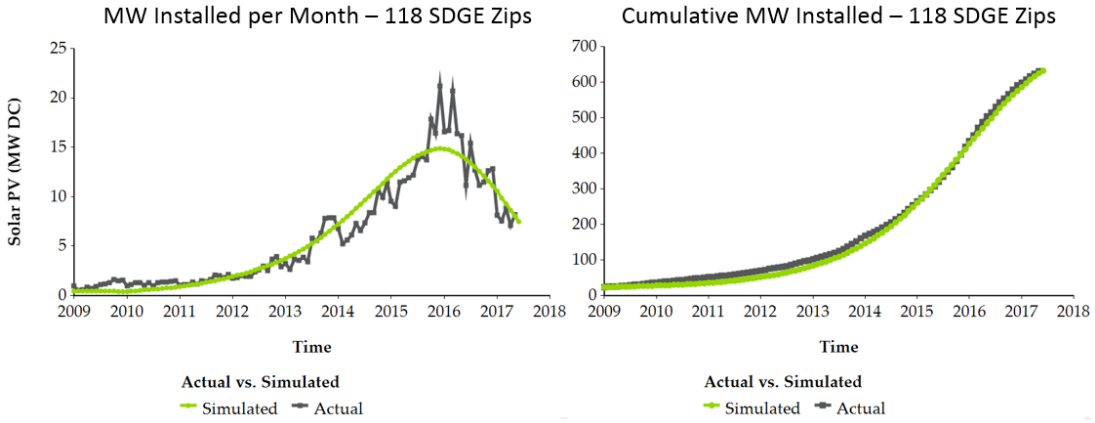
The last key customer attribute identified in the ML analysis was the percentage of owner-occupied housing units. The project team accounted for this parameter more directly by calculating an eligible homes value, which only included owner-occupied homes and the calculated fraction of customers (by ZIP code) whose credit score exceeded 680, a typical value assumed to qualify for a solar PV lease or PPA.

To estimate the values of the above coefficients, the project team conducted a logit regression analysis by transforming the long-run market share equation into a linear function, permitting linear regression of each of the coefficients. The team calculated the market share input for the regression by applying a scaling factor against the actual market share that was a function of the best-fit diffusion parameters from the prior step in the analysis.

After calculating the coefficients of the logit market share equation, the project team employed a nonlinear optimization to re-calculate the global (i.e., constant across ZIP codes) diffusion coefficients, which the team expected to be different from the prior analysis because the calculation of long-run market share has changed with the additional coefficients in the calculation.

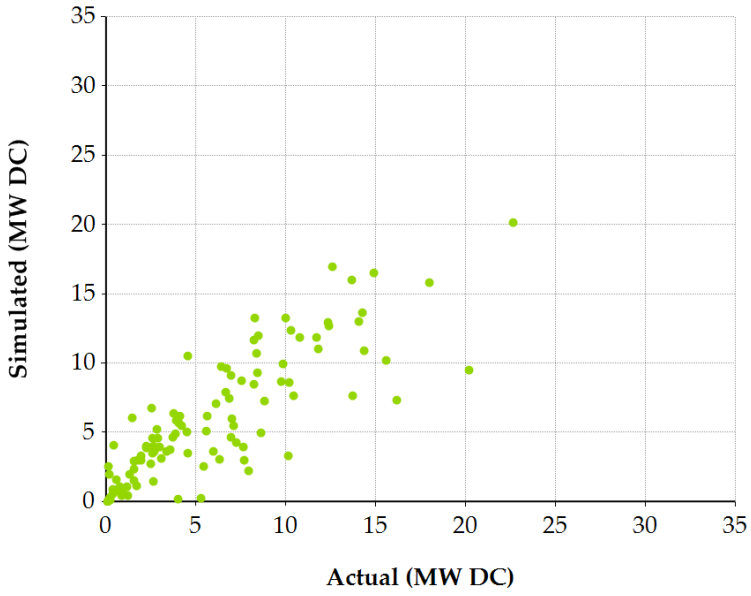
The analysis results again show a strong fit of incremental and cumulative adoption from 2009 through the first quarter of 2017, as illustrated below in Figure 13.

Figure 13. Model Simulated vs. Actual PV Installed – Adding Customer Attribute Coefficients



The project team again plotted the simulated versus actual cumulative adoption from January 2009 through March 2017, as seen in Figure 14. Comparing this figure with Figure 13, a better correlation emerged between the simulated and actual adoption forecasted at the ZIP code level, yet considerable variance between simulated and actual still exists. Adding the customer attribute coefficients reduces the value of the objective function[8] for the parameter fitting optimization by roughly 33%.

Figure 14. Simulated Actual Cumulative Adoption (March 2017) for all 118 SDG&E ZIP Codes – Adding Customer Attribute Coefficients



[8] The objective function in the parameter fit is the sum of the absolute value of the difference between simulated and actual cumulative adoption of each ZIP code summed over every month of the simulation and over all ZIP codes.

2.1.4.3 Granular Adoption Analysis with ZIP Code-Specific Parameters

While adding customer attributes to the long-run market share calculation somewhat improved the fit at the ZIP code level, sufficient variance still exists to warrant further exploration of model parameters. When almost every ZIP code analyzed appears to be past the inflection point of adoption, adding another parameter to the market share calculation specific to each ZIP code should provide an even better fit at the ZIP code level. Addition of such granular parameters can be risky when little information is known about the S-curve adoption profile and can result in poor long-run market share forecasts when early in the adoption curve. In these cases, a danger of overfitting the model exists, resulting in a false sense of security regarding the forecasting accuracy based on past data fits. However, once past the inflection point of adoption, this risk diminishes greatly, as the shape of the S-curve and of the incremental adoption data indicates the ultimate market saturation value.

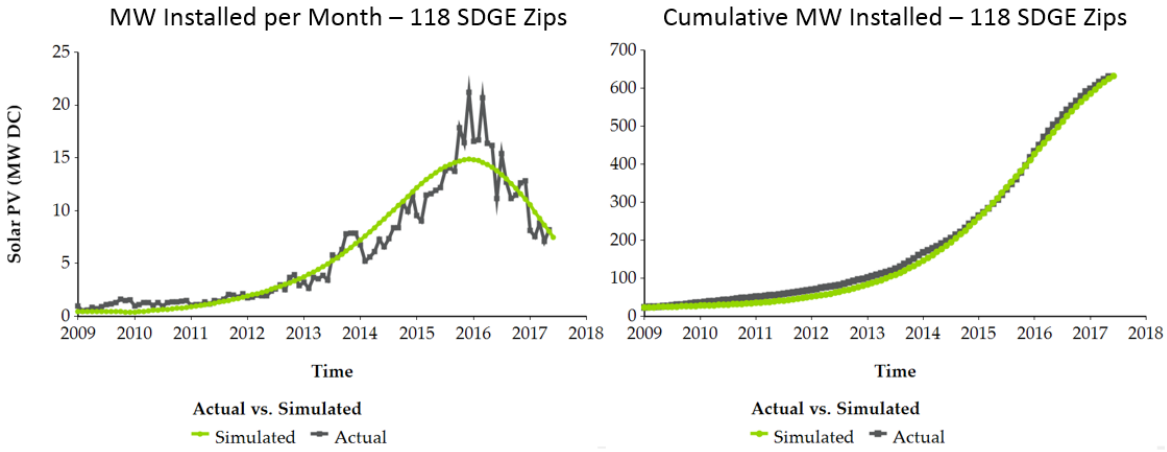
In theory, with sufficient historical data for each ZIP code, one could fit all the parameters of the model to the data, resulting in ZIP code-specific parameters for both the long-run market share calculations and for the coefficients governing the rate of diffusion (the shape of the S-curve). However, these optimizations take time to perform, especially when conducting them across 118 ZIP codes in SDG&E's service territory. Additionally, ensuring a good fit of the data is sometimes an iterative process, involving the addition of constraints or tightening the upper and lower bounds of the parameter estimates to ensure good parameter estimates. The reason for this is the inherent limitations of a nonlinear optimization, which uses gradient hill climbing techniques, which are prone to finding local optima as opposed to global optima. Considering the scope limitations of this project, the project team chose only to fit a single ZIP code-specific parameter, adding an a0 term to the calculation of long-run market (refer to Equation 2 for the original formula). For simplicity of code modification and to facilitate comparison of results, the project team added an a0 term to the original expression for the α term in lieu of replacing the entire α term with only a ZIP code-specific parameter. The results of the fit are the same in either case since the optimization of an alternate a0 term would effectively roll into it all the other terms in the α expression in a single, modified coefficient. In other words, the customer-specific attributes calculated in the prior step become redundant when one adds a ZIP code-specific parameter to the calculation of long-run market share.

Equation 4. Disaggregation of Logit Coefficient

$$\alpha_{ZipCode} = a0_{ZipCode} + a1_{Climate Zone} + a2 * attribute2_{ZipCode} + a3 * attribute3_{ZipCode} + a4 * attribute4_{ZipCode} + a5 * attribute5_{ZipCode}$$

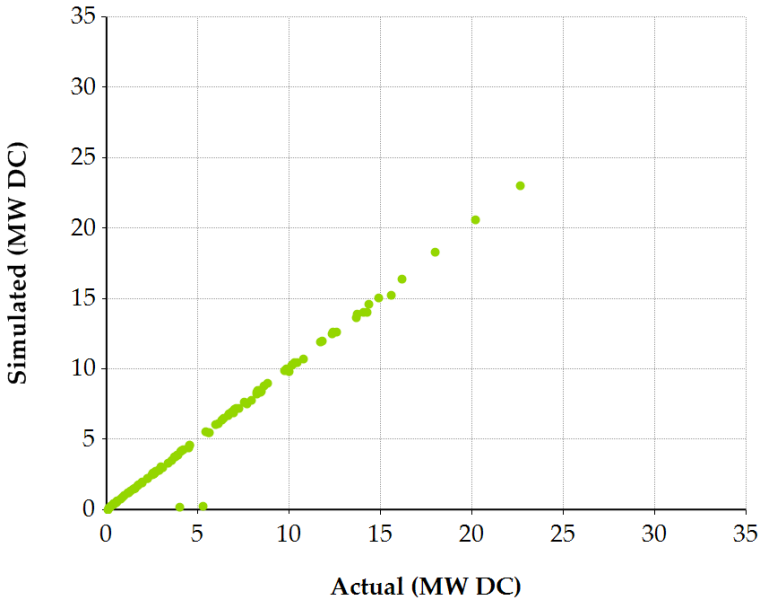
Again, the results demonstrated an excellent fit of both incremental adoption and cumulative adoption when compared with the historical data, as shown in Figure 15.

Figure 15. Model Simulated vs. Actual PV Installed



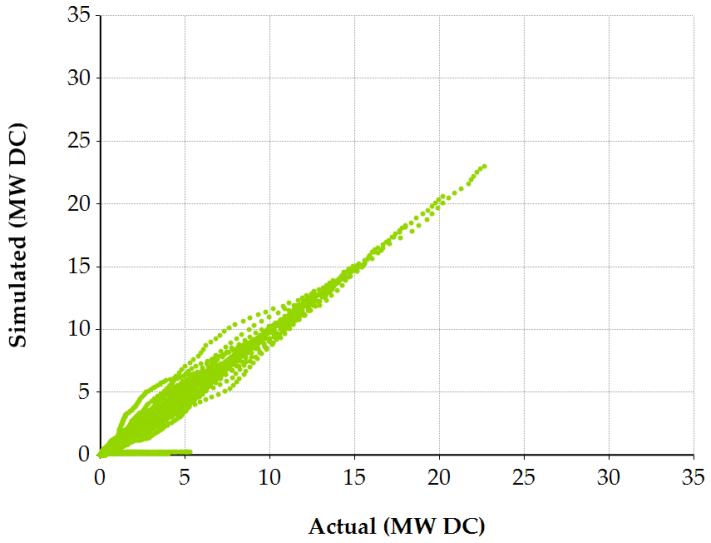
Looking at the results at the ZIP code level, however, there is a much better fit. Simulated cumulative adoption from 2009 through March 2017 aligns closely with actual historical adoption for nearly every ZIP code analyzed. These results can be compared with those presented in Figure 11 and Figure 13.

Figure 16. Simulated vs. Actual Cumulative Adoption (in March 2017) for all 118 SDG&E ZIP Codes – Adding ZIP Code-Specific Coefficients



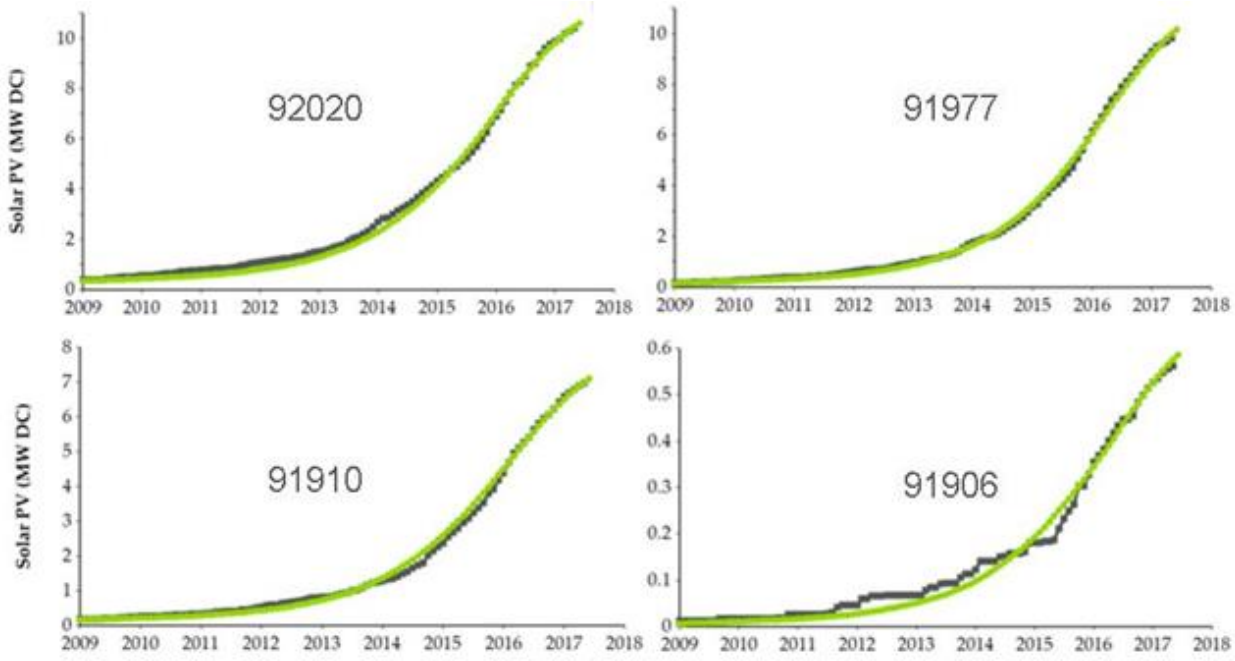
Due to the construct of the optimization and the constraints imposed, the fit is best at the final timestep of the simulation. Greater variance exists between simulated and actual cumulative adoption at each timestep in the simulation. In Figure 17, where each data point represents an individual ZIP code in each month of the simulation from 2009 through 2017, additional variance is observed when inspecting the curve fits at all time periods of the simulation; however, in general, the simulated cumulative adoption fits the actual cumulative adoption very well.

Figure 17. Monthly Simulated vs. Actual Cumulative Adoption (January 2009 – March 2017) for all 118 SDG&E ZIP Codes – Adding ZIP Code-Specific Coefficients



Another way to inspect the data is to look at the simulated versus actual adoption over time for individual ZIP codes. Though showing this result for all 118 ZIP codes would be cumbersome, Figure 18 shows the goodness of fit of the simulated adoption with actual cumulative adoption for four typical ZIP codes. As shown in the figure, the simulated adoption aligns quite well with actual adoption over time even at the ZIP code level, though actual adoption data tends to be a bit less smooth when viewing at the individual ZIP code level (particularly for ZIP codes with smaller installation quantities) as opposed to summed over many ZIP codes. Though not all ZIP codes align as well as those illustrated below, these results are typical and reasonably consistent at the ZIP code level, though outliers do exist.

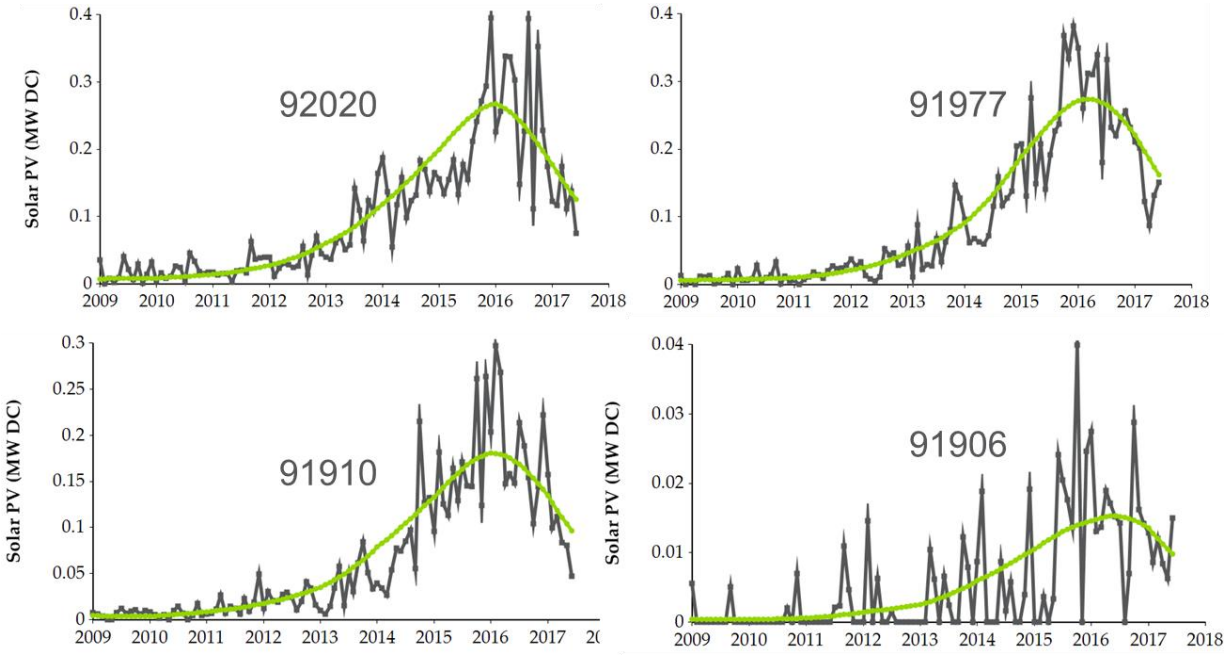
Figure 18. Simulated vs. Actual Cumulative Adoption over Time – ZIP Codes 92020, 91977, 91910, and 91906



Though each of the above adoption profiles results in a different long-run market share, the shape of the adoption (i.e., the shape of the S-curve, or the rate at which the long-run market share is approached) is reasonably consistent across ZIP codes. Again, outliers do exist, but for the most part, assuming a constant set of diffusion coefficients (i.e., *p* and *q* in the classical Bass model) across the service territory is a reasonable approximation.

Figure 19 shows the incremental adoption per month for each of the above four ZIP codes. Whenever one views incremental data as opposed to cumulative data, the results are noisier, owing to the smoothing effect of integrating—or cumulating—data over time. That said, the project team still observes a strong correlation between simulated and actual adoption at the ZIP code level, even when comparing incremental adoption data. It should also be apparent that the characteristic rise and fall of incremental adoption, manifested as a somewhat bell-shaped incremental adoption curve, is visible at the ZIP code level, not just in aggregate across ZIP codes.

Figure 19. Simulated vs. Actual Monthly Incremental Adoption over Time – ZIP Codes 92020, 91977, 91910, and 91906



2.2 Disadvantaged Communities Results and Analysis

The following sections detail the results the DAC demonstration activity.

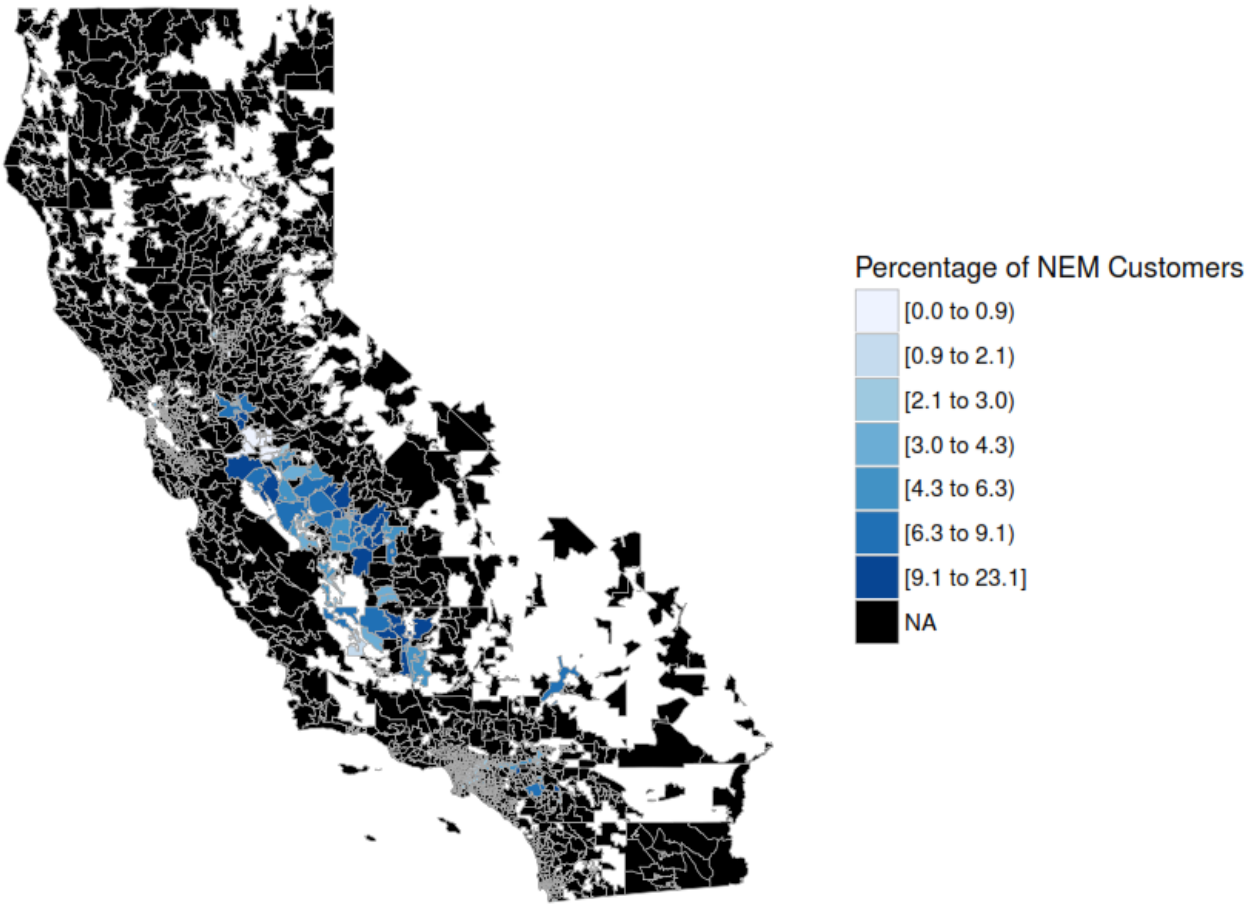
2.2.1 Customer Attributes Driving Adoption

The project team conducted ML modeling to identify the most important attributes driving adoption at the ZIP code level, for DAC-designated ZIP codes. This analysis included filtering the compiled analysis data as described in Section 2.1.2 to only include those with DAC indicators and available attributes from the American Community Survey at the ZIP code level, and fitting the ML models.

2.2.2 Data Compilation

After filtering the California IOU territory ZIP code data, the resulting dataset included 171 out of 255 ZIP codes designated as DAC ZIP codes due to NEM data only being available in IOU territories and American Community Survey census data limitations. The NEM installation penetration for the DAC ZIP codes is shown in Figure 20 in shades of blue from light (lowest percentage) to dark (highest percentage), with black signifying non-DAC ZIP codes or outside of the IOU service territories, and white signifying Census Bureau ZIP Code level attribute data were not available.

Figure 20. NEM Installation Percentage, by DAC ZIP Codes



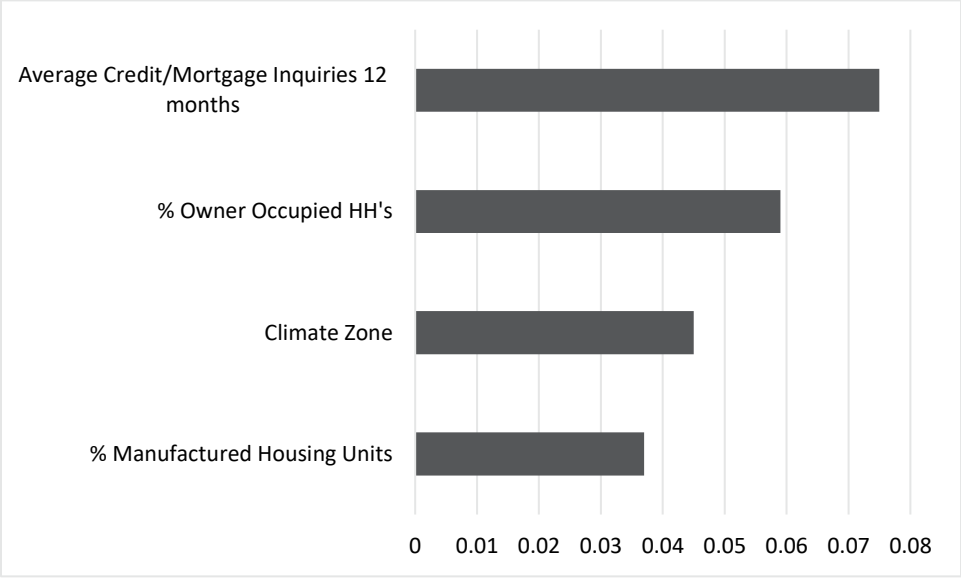
2.2.3 Machine Learning Models

The project team fit a Conditional Inference Tree (CTree) model to the DAC-filtered set of California ZIP codes for estimating the NEM percentage, and found that it outperformed the random forest model in terms of variance explained, when fit on the full set of DAC ZIP codes and the full set of candidate predictor variables. CTree models account for statistical distributional properties of candidate predictor variables when considering non-parametric splits, as opposed to recursive univariate splits used for the individual CART trees in a random forest. [25]

The CTree model isolated four attributes most strongly linked to the NEM percentage in DAC ZIP codes, as shown below in Figure 21.

The project team fed the attributes from the CTree model into a random forest ensemble model to quantify the variable importance and to test whether the importance order observed was consistent. Overall, the random forest model explained 53 percent of the variance in NEM installation percentage for the DAC ZIP codes. The attribute importance order from the CTree was also observed in the random forest model, shown below.

Figure 21. Random Forest Attribute Variable Importance – DAC ZIP Codes



Three of these four key explanatory attributes were also significant for the overall NEM Installation ML analysis, with only the percentage of manufactured housing units found to be significantly linked to predicting NEM installation within the DAC filtered list of ZIP codes.

The figure below depicts the random forest error distribution for estimating the proportion of households with NEM installs for solar PV system installations—i.e., the error of the ML model to the actual DG Stats data, in the DAC ZIP codes. The black curve shows the estimation error distribution, centered on zero, and slightly left skewed, implying the random forest models had a slight tendency to under-state the NEM installation percentage.

Figure 22. Random Forest Estimation Error Distribution for Percentage of Households with NEM Installations

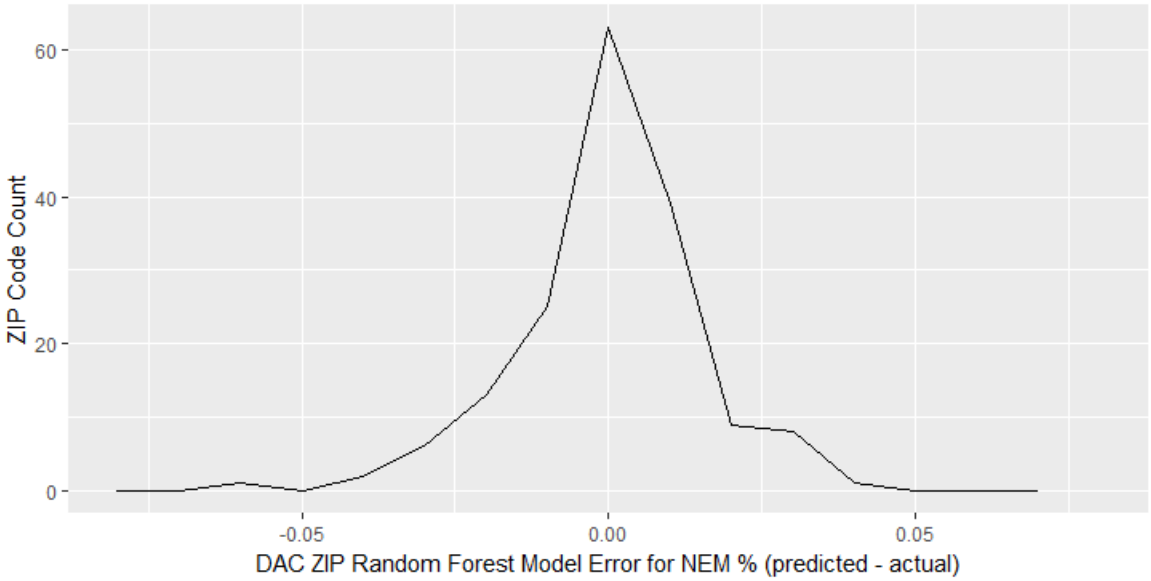


Table 6 details the numerical summary of the error distribution shown in Figure 22. Approximately 50 percent of the DAC ZIP codes were predicted within 0.1% of their true NEM installation percentage.

Table 6. Random Forest Estimation Error Quartiles and Mean Percentage of Households with NEM Installs

DAC ZIP Codes	Min	1st Quartile	Median	Mean	3rd Quartile	Max
SDG&E	-5.7%	-0.1%	0.0%	0.0%	0.1%	4.4%

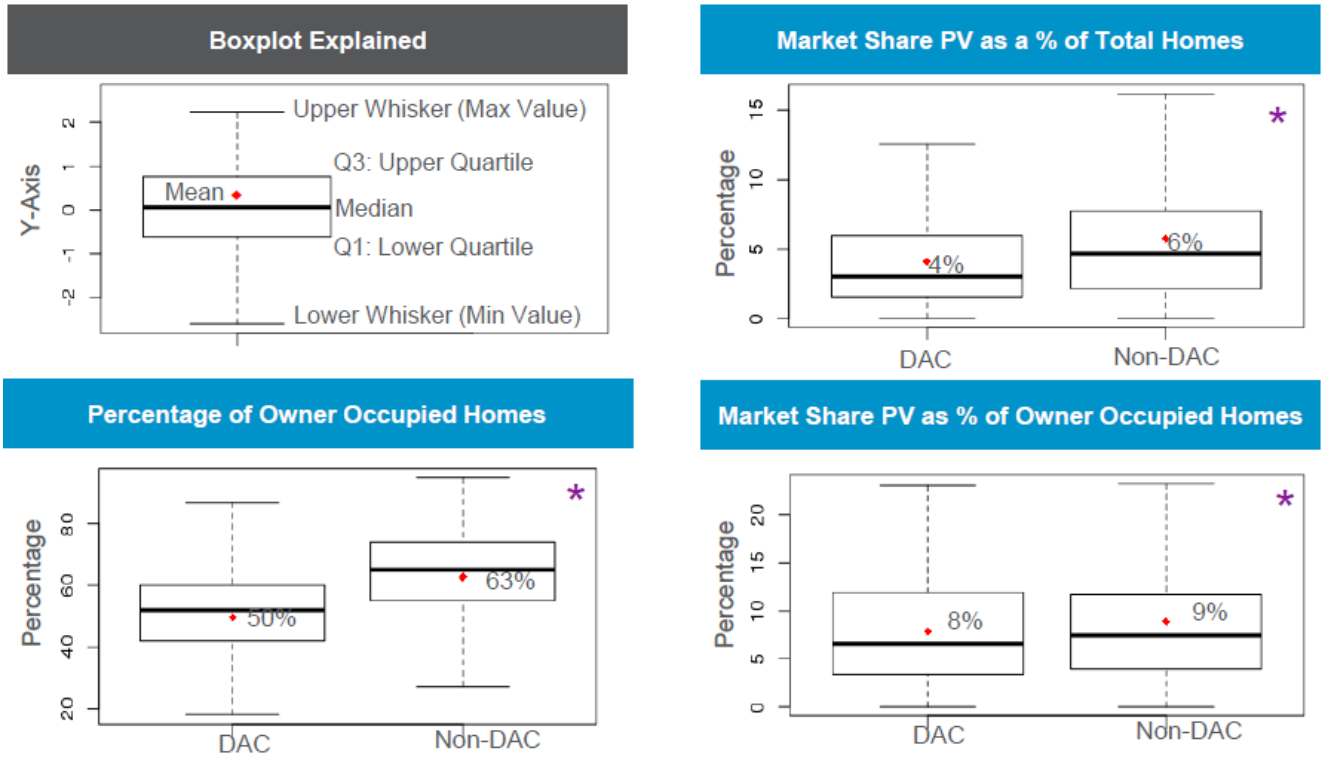
2.2.4 Statistical Analysis Results

Figure 23 and Figure 24 depict the statistical analysis results as box and whisker plots. As illustrated below, owner occupancy is a key attribute explaining the difference in solar PV market share between DAC and non-DAC ZIP codes. The percentage of owner occupied homes is 63% for non-DAC ZIP codes. This is statistically different from the average home ownership in DAC ZIP codes of 50%.[9] The solar PV market share as a percentage of total homes is 6% for non-DAC ZIP codes, versus DAC ZIP codes where solar PV market share is 4%, signifying a statistically different mean value.

When analyzing the solar PV market share as a percent of owner occupied homes (the bottom right graph in Figure 23), the difference in average market share across DAC and non-DAC ZIP codes is still statistically significant; however, the distributions are more similar when only evaluating owner occupied homes.

[9] Statistical significance is measured at the 90% confidence interval.

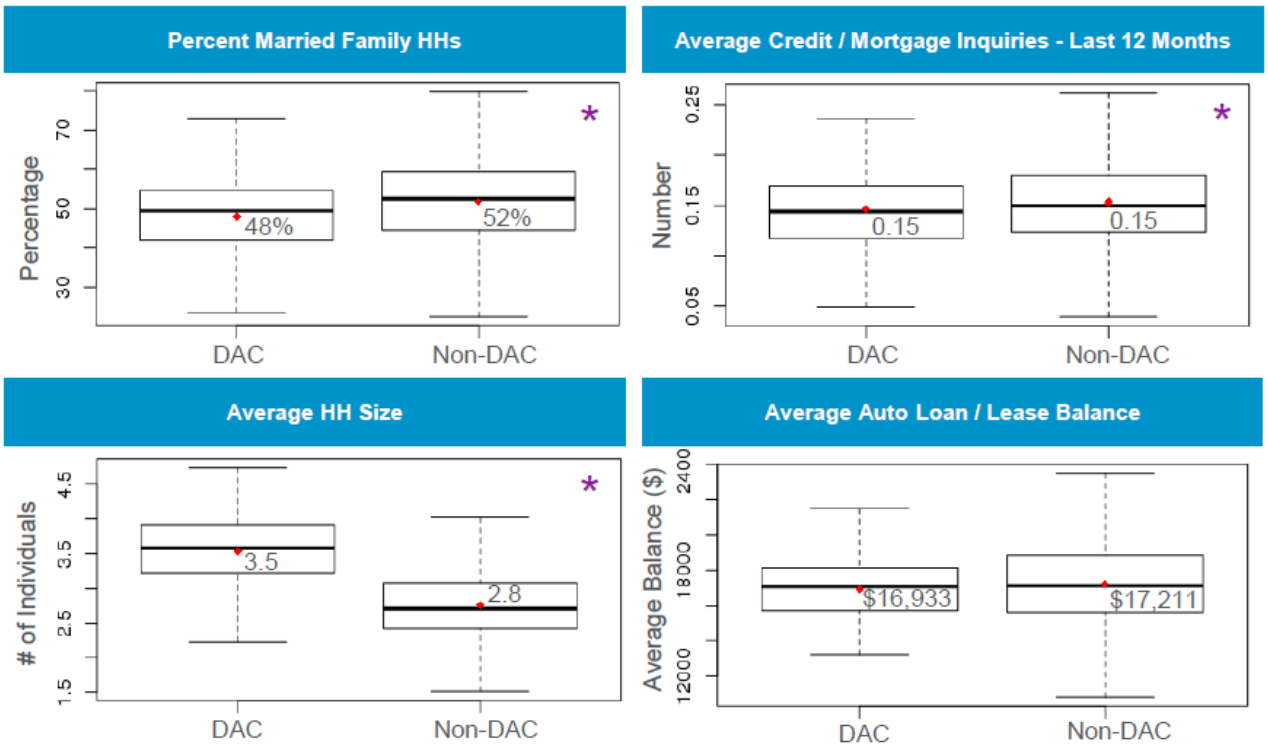
Figure 23. Percentage of Owner Occupied Homes and Market Share Attributes



**Indicates that the means of the attribute between DAC and Non-DAC ZIPs are statistically different at the 90% confidence interval.*

As illustrated in Figure 24, the remaining four attributes are more similarly distributed between DAC and non-DAC and although statistically significant do not practically explain differences in adoption. For these parameters, a higher value correlates with higher adoption levels for all attributes other than household size. Yet the positive correlation of the other parameters, particularly homeownership, overwhelms the solar PV adoption results.

Figure 24. Remaining Four Attributes

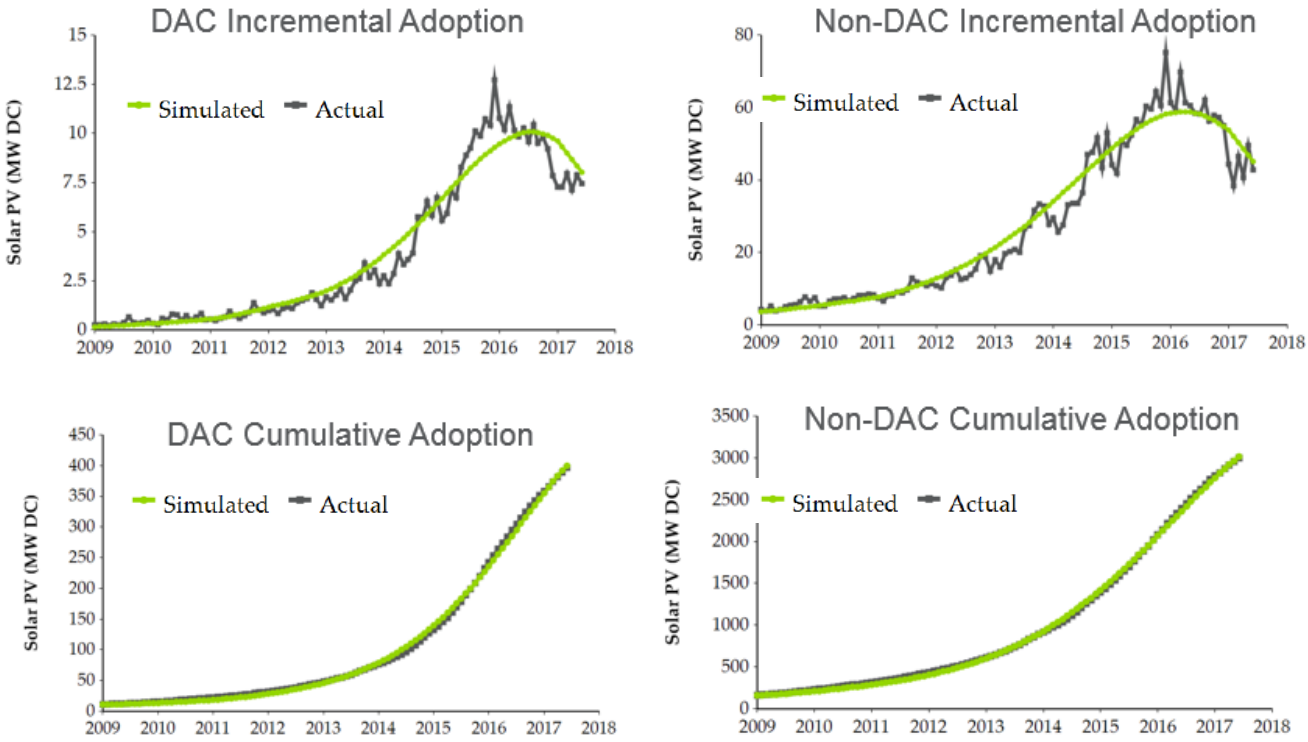


**Indicates that the means of the attribute between DAC and Non-DAC ZIPs are statistically different at the 90% confidence interval; however, this does not signify that these attributes are drivers of differences between DAC and non-DAC.*

2.2.5 DAC vs. Non-DAC Adoption Analysis

Comparing the model simulated with actual adoption data for DAC Zip codes and non-DAC Zip codes, one finds that both DAC and non-DAC data fit simulated results quite well, as illustrated below in Figure 25. Both categories again show the characteristic rise and fall of incremental adoption, resulting in S-shaped cumulative adoption that is similarly shaped across both categories. Differences in the rates of adoption or shape of the S-curve appear minor; however, one may notice a somewhat greater curvature in DAC incremental adoption in the 2015-2016 timeframe, possibly a result of the timing of the extension of the California Single Family Affordable Solar Housing (SASH) and Multi-Family Affordable Solar Housing (MASH) incentives. [26] As such, the non-DAC simulated results tend to fit the data somewhat more closely, though both sets of simulated results compare well with actual historical adoption.

Figure 25. Comparison of DAC and non-DAC Adoption: Simulated vs. Actual



2.3 Assumptions

The team made some specific assumptions:

- Considered only credit score eligibility and homeownership in calculating suitability for installation, as opposed to further refining technical suitability of homes for solar PV (e.g., fraction of homes in each ZIP code with acceptable orientation, shading, etc.).
- Utilized constant pricing undercut of leases/PPAs relative to prevailing electricity rates, as opposed to calculating or estimating pricing undercut differences over time.
- Aggregated analysis to the ZIP code level rather than the customer unit level.
- Defined a DAC ZIP code as a ZIP code where >50% of the population lives in census tracts with a CalEnviroScreen 3.0 score > 75%.
- Did not use ZIP Codes without a direct 1:1 map with ZIP Code Tabulation Areas (ZCTAs) in the DAC analysis. The excluded ZIP codes were found in small population density areas within the IOU service territories, or had very few residential customers. The Census Bureau does not publish demographics data for these ZIP codes to protect confidentiality.

3. Key Findings

The following section presents the key project findings that were identified by the project team. The strength of aggregate and ZIP-code-level back-casts suggest that causal models can be used to forecast residential rooftop PV adoption moving forward with a reasonable degree of accuracy, even when the analysis is spatially disaggregated. Such methods could support integrated resource planning and a better understanding of likely solar PV installation location, facilitating transmission and distribution-level planning and analysis.

As indicated by the demonstration results, the key findings of this project can be summarized as follows:

- Causal models, when appropriately calibrated, can explain historical adoption patterns well.
- The Bass diffusion construct employed in the project team's BDCM can replicate historical adoption patterns for both cumulative and incremental solar PV adoption.
- Preliminary results suggest that the adoption of residential solar PV in California and in the SDG&E service territory is past the inflection point in the characteristic S-curve of adoption.
- ML techniques can help to explain historical adoption patterns and can reduce the variance between simulated and actual adoption when analyzed at a granular level. Importantly, these techniques confirmed existing assumptions (e.g. climate zone) and introduced new considerations (e.g. married family households).
- The parameterization of the most important attributes through discrete choice modeling— are best suited to understanding drivers of individual customer adoption, which was beyond the scope of this demonstration
- The rate of diffusion (i.e., the shape of the S-curve) is reasonably consistent spatially; the long-run market share appears to differ more substantially when analyzed at a granular level (e.g., at the ZIP code level).
- ZIP code-level forecasts seem possible with reasonable accuracy, particularly when ZIP code-level adoption appears to be past the inflection point of S-curve adoption.
- Market share solar PV as a percent of total homes is about 50% higher in non-DAC ZIP codes (6% vs. 4% market share).
- Owner occupancy is a key attribute in explaining the differences between DAC and non-DAC solar PV market adoption.
    - When comparing PV adoption in only owner-occupied homes, market share solar PV is very similar between non-DAC and DAC ZIP codes (9% vs. 8%).
    - Differences in other attributes do not drive strong variance in solar PV market share across non-DAC and DAC ZIP codes.
- Statewide analysis of Non-DAC and DAC ZIP codes generates close fit between simulated and actual. Although DAC ZIP codes have a lower penetration as a percentage of total homes, market share is closer as a percentage of owner occupied homes.

The project team views the outcome of this pre-commercial demonstration as a success because the demonstration activity results verify the capabilities of the methodology that was demonstrated. The ability of the methodology to forecast spatially was even more certain than expected at the outset of the project due to the DGStats installation data suggesting the market may be past the inflection point of the S-curve. As DER penetration continues to grow, this project outlines how SDG&E and other utilities can leverage the methodology framework to help forecast the adoption of other DER in their jurisdictions.

4. Recommendations and Next Steps

The project has demonstrated a set of methods and tools that could be used to estimate the propensity for customer adoption of a DER technology such as photovoltaics. The project team recommends that SDG&E not commercially adopt these methods and tools at this juncture, without more foundational work being done first. Based on the pre-commercial demonstration results and findings of this project, the following actions by SDG&E or other stakeholders are recommended as steps toward prospective commercial adoption of the demonstrated methods and tools.

- Improve SDG&E's existing zip-code based Bass diffusion technique with refinements for the long-run market share parameters based on significant customer attributes.
- Improve certain model inputs (e.g., historical PPA prices, kilowatt-hour production, technical suitability due to shading and orientation, price sensitivity, and correlation between homeownership and credit scores) and conduct additional research to determine whether the residential solar PV market in SDG&E territory is approaching saturation.
    - Refine estimate of suitable buildings in the SDG&E service territory
    - Refine price sensitivity to forecast future adoption under different price/policy scenarios
    - Add parameters at the ZIP code level (e.g., one or two of the diffusion parameters) to provide greater forecast accuracy at that level of granularity
- Leverage the same or equivalent methodology to evaluate solar PV adoption for other specific segments of interest and potentially individual customer analysis, including but not limited to the following groups:
    - Commercial and industrial customers
    - Low-income customers building on the analysis done on DAC
    - Customers on distribution feeders that are capacity constrained or at risk for reverse power flow during peak PV generation hours
- Adapt the methodology for use in forecasting adoption of other DER (e.g., solar + storage, storage, EVs) and conduct demonstrations in these areas.
- Consider utilizing a customer discrete choice survey approach to facilitate independent estimation of both the long-run market share parameters and the Bass diffusion coefficients.

5. Metrics and Value Proposition

5.1 Metrics

The commercial adoption methodologies and tools for estimating propensity for customer adoption of photovoltaics will be impacted by the following metrics.

- Potential energy and cost savings
    - Avoided customer energy use (kWh saved) – The use of tools to estimate customer adoption of PV would lead to understanding the contribution of electric load from PV systems, which in turn will provide the customers with reduced energy usage and economic savings.
    - Avoided procurement and generation costs – Accurate estimation of customer PV adoption rates would enable utilities to estimate the avoided cost to procure energy from sources that might be inefficient or contribute to environmental pollution.
- Environmental benefits
    - GHG emissions reductions – Adoption of PV would lead to reduced emissions from fossil fuel based sources which would have to be used in absence of renewable resources like PV.

5.2 Value Proposition

The purpose of EPIC funding is to support investments in R&D projects that benefit the electricity customers of SDG&E, PG&E, and SCE. The primary principles of EPIC are to invest in technologies and approaches that provide benefits to electric ratepayers by promoting greater reliability, lower costs, and increased safety. This EPIC project contributes to these primary principles in the following ways:

- Greater Reliability: More accurate DER forecasting techniques will be required as these technologies have a greater impact on SDG&E's distribution system. It has become evident through circuit load data that residential PV is now playing a role in daily load shapes. To ensure the system is properly designed for future needs, PV adoption forecasts must be carefully analyzed to anticipate future electric system requirements and reduce the risk of outages.
- Lower costs: PV adoption will likely have a direct impact on the type and location of distribution system, and possibly, transmission system upgrades. Applying the most appropriate resources at the most beneficial locations will inherently keep costs lower than the alternatives. Improved forecasting methods should enable the allocation of those resources to be applied in the most appropriate way.
- Higher consumer satisfaction: More accurate DER forecasting can improve consumers' contribution through demand response management in the operation of a utility power system by reducing or shifting their electricity usage during peak periods in response to time-based rates or other forms of financial incentives.

6. Technology Transfer Plan

6.1 SDG&E Technology Transfer Plans

A primary benefit of the EPIC program is the technology and knowledge sharing that occurs both internally within SDG&E and across the industry. To facilitate this knowledge sharing, SDG&E will share the results of this project by announcing the availability of this report to industry stakeholders on its EPIC website, by submitting papers to technical journals and conferences, and by presentations in EPIC and other industry workshops and forums. The results will also be shared internally through presentations to internal stakeholders.

6.2 Adaptability to Other Utilities and Industry

All successful product rollouts tend to follow an S-shaped adoption curve, and solar PV is no exception. The further a technology progresses along the S-curve, the more accurate adoption forecasts become. The project approach is readily adaptable to all other IOUs in California, with adoption patterns holding across most ZIP codes in the state. This approach to forecasting adoption can be utilized by other utilities for integrated resource planning and forecasting transmission and distribution needs. While this demonstration project leveraged public information, if the project team had customer-specific data, the granularity of the analysis could have been refined, possibly to the feeder and substation level under various scenarios and rate structures.

7. References

[1] Wikipedia. “ZIP Code.” Internet: https://en.wikipedia.org/wiki/ZIP_Code, Dec. 5, 2017.

[2] D. McFadden. “Economic Choices,” presented at the Prize Lecture, Stockholm, Sweden, 2000.

[3] M.E. Ben-Akiva, S.R. Lerman. Discrete Choice Analysis: Theory and Application to Travel Demand. Cambridge, MA: The MIT Press, 2006.

[4] L. O'Keeffe. “A Choice Experiment Survey Analysis of Public Preferences for Renewable Energy in the United States,” Journal of Environmental and Resource Economics at Colby, vol. 01, 2014.

[5] F.M. Bass. “A New Product Growth for Model Consumer Durables.” Management Science, vol. 15, pp. 215-227, Jan. 1969.

[6] F.M. Bass. “Comments on ‘A New Product Growth for Model Consumer Durables: The Bass Model.’” Management Science, vol. 50, pp. 1833–1840, Dec. 2004.

[7] J.D. Sterman. “The Bass Diffusion Model,” in Business Dynamics: Systems Thinking and Modeling for a Complex World, S. Isenberg, New York: McGraw-Hill, 2000, pp. 332.

[8] V. Mahajan, E. Muller, and Y. Wind. “Diffusion Models, Managerial Applications and Software,” in New-Product Diffusion Models. New York: Springer, 2000, pp. 295-310.

[9] B. Sigrin, M. Gleason, R. Preus, I. Baring-Gould, and R. Margolis. “The Distributed Generation Market Demand Model (dGen): Documentation.” Internet: https://www.nrel.gov/docs/fy16osti/65231.pdf. Feb. 2016.

[10] A. Agarwal. “A Model for Residential Adoption of Photovoltaic Systems.” M.S. thesis, California Institute of Technology, California, 2015.

[11] L. Breiman. Classification and Regression Trees. California: Wadsworth & Brooks/Cole Advanced Books & Software, 1984.

[12] L. Breiman. Classification and Regression Trees. California: Wadsworth & Brooks/Cole Advanced Books & Software, 1984.

[13] W.L. Loh. “Fifty Years of Classification and Regression Trees.” International Statistical Review, vol. 84, pp. 329-348, 2014.

[14] A Liaw and M. Wiener. “Classification and Regression by Random Forest.” R News, vol. 2/3, pp. 18-22, Dec. 2002.

[15] S. Athey and I. Guido. “Machine Learning Methods for Causal Effects.” Internet: www.nasonline.org/programs/sackler-colloquia/documents/athey.pdf. 2015.

[16] P. Bajari, D. Nekipelov, S. Ryan, and M. Yang. “Demand Estimation with Machine Learning and Model Combination.” National Bureau of Economic Research, 2015.

[17] US Environmental Protection Agency. “Understanding Third-Party Ownership Financing Structures for Renewable Energy.” Internet: www.epa.gov/repowertoolbox/understanding-third-party-ownership-financing-structures-renewable-energy.

[18] F. Stermole and J. Stermole. Economic Evaluation and Investment Decision Methods. Investment Evaluation Corporation, 2012.

[19] Navigant Consulting, Inc., Solar Project Return Analysis for Third Party Owned Solar Systems. 2016.

[20] U. Benzion and J. Yagil. “Decisions in Financial Economics: An Experimental Study of Discount Rates,” Advances in Financial Economics, M. Hirschey, J. Kose, A.K. Makhija, Eds. United Kingdom: Emerald Group Publishing, 2002, pp. 19-40.

[21] California Distributed Generation Statistics. “NEM Currently Interconnected Data Set.” Internet: http://www.californiadgstats.ca.gov/downloads/, Sept. 30, 2017.

[22] California Energy Commission. “Electric Program Investment Charge 2016 Annual Report.” Internet: http://www.energy.ca.gov/2017publications/CEC-500-2017-015/CEC-500-2017-015-CMF.pdf, April 2017.

[23] California Distributed Generation Statistics. “Currently Interconnected Data Set.” Internet: http://www.californiadgstats.ca.gov, 2017.

[24] Navigant Consulting, Inc., “Solar Project Return Analysis for Third Party Owned Solar Systems.” Proc. UniSource Electric Rate Case, 2016.

[25] T. Hothorn, K. Hornik, and A. Zeileis. “ctree: Conditional Inference Trees.” Internet: https://cran.r-project.org/web/packages/partykit/vignettes/ctree.pdf, 2015.

[26] “Decision Extending the Multifamily Affordable Solar Housing and Single Family Affordable Solar Homes Programs within the California Solar Initiative.” Internet: http://docs.cpuc.ca.gov/PublishedDocs/Published/G000/M145/K938/145938475.PDF, Jan. 29, 2015.