Screening Solar PV Adopters and Non-adopters: An Application of Machine Learning Methods

Changgui Dong†, Benjamin Sigrin‡

† Renmin University of China

‡ National Renewable Energy Laboratory

November 13, 2017

Presentation at the 35th USAEE/IAEE North American Conference

RENMIN UNIVERSITY OF CHINA

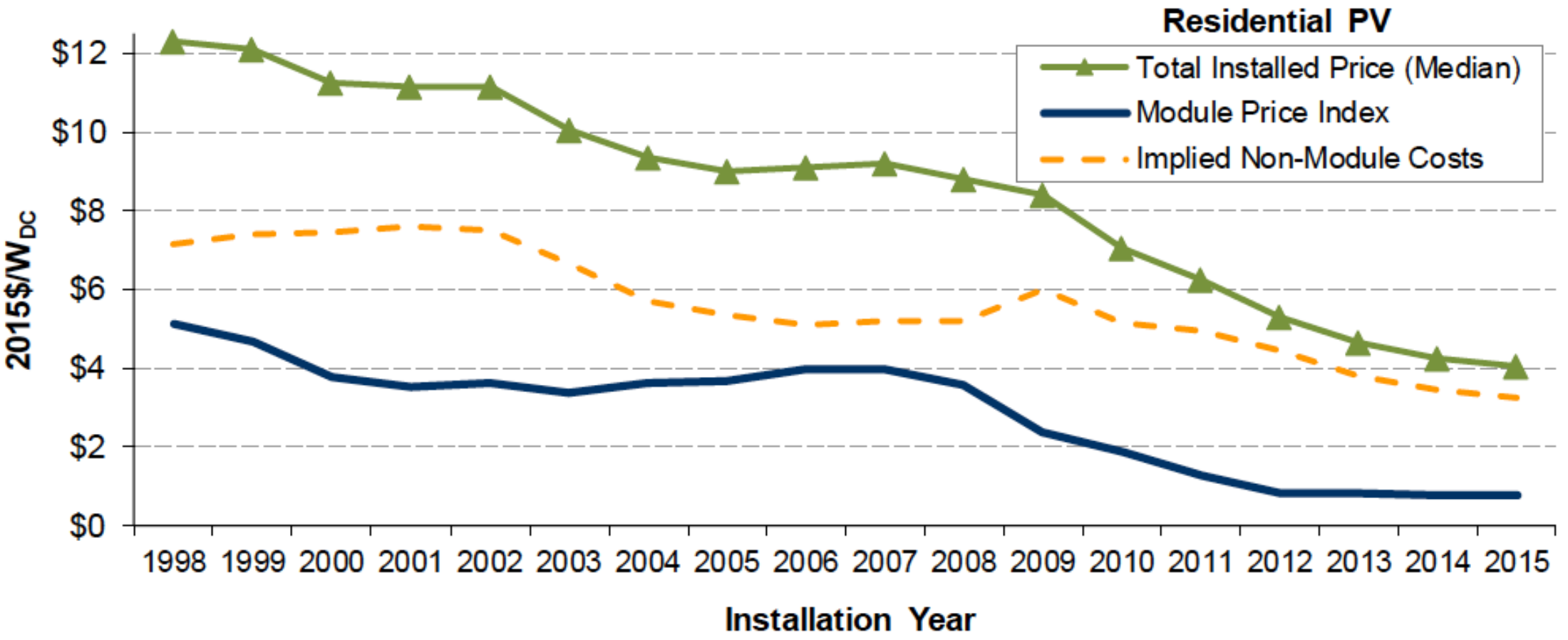
NATIONAL RENEWABLE ENERGY LABORATORY

Research Question

Can “visible” variables be used to better screen for PV adoption likelihood and thus reduce PV companies’ customer acquisition costs?

- Improved screening is important, not only for generating more PV adoptions and mitigating climate change, but also for advancing the usage of state-of-the-art techniques, such as machine learning, in research methods.

U.S. Residential Solar Cost Structure

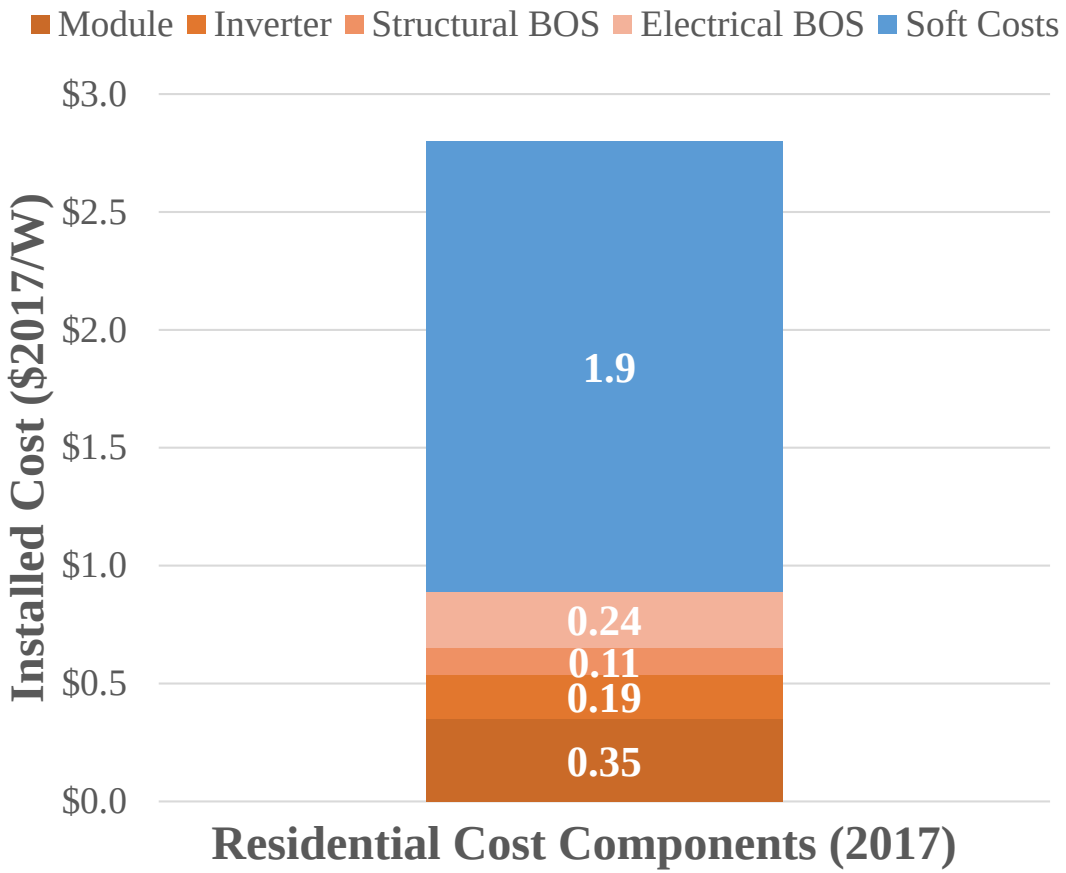


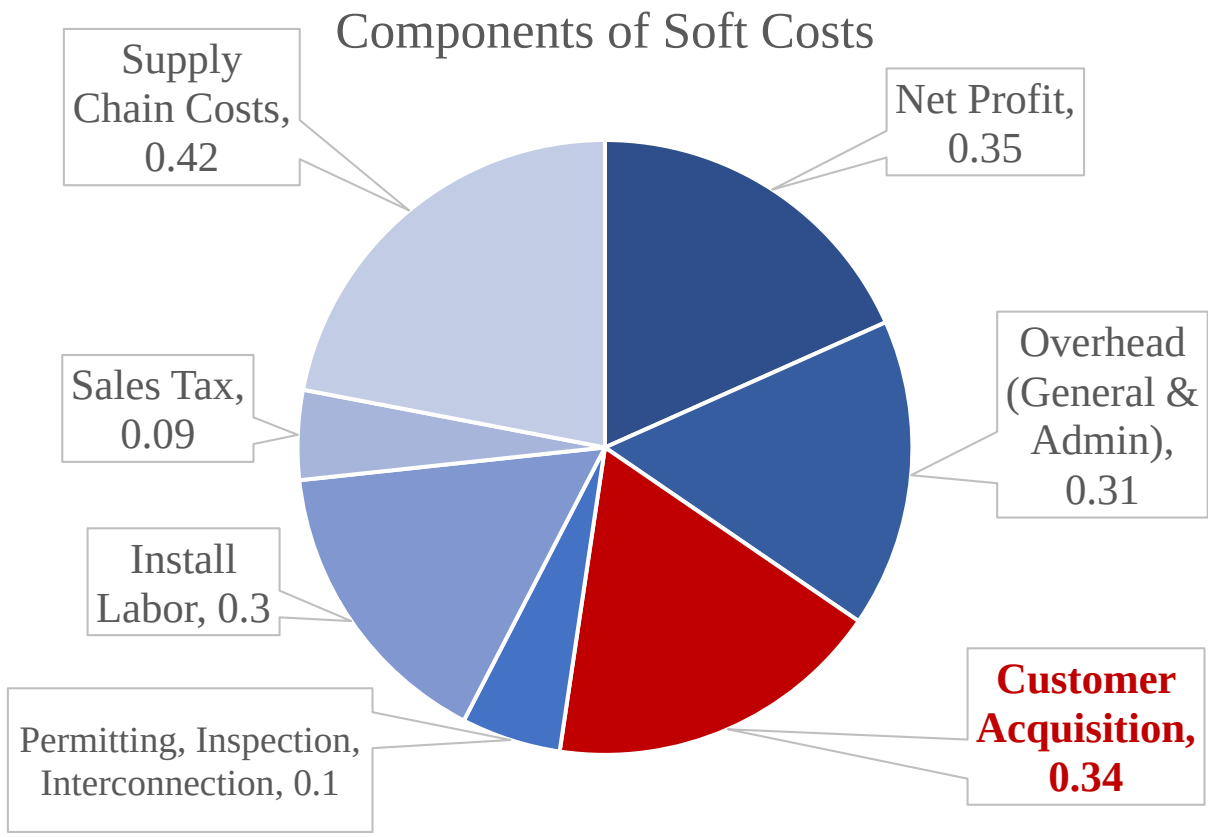
Notes: The Module Price Index is the U.S. module price index published by SPV Market Research (Mints 2015). Implied Non-Module Costs are calculated as the Total Installed Price minus the Module Price Index, and therefore include installer profit margin.

Source: Barbose et al., 2016. Tracking the Sun IX

2017 U.S. Residential Solar Cost Structure

- Residential solar installed costs were $2.8/W in 2017, 64% of which are soft costs.
- Customer acquisition costs ($0.34/W) were equal to the cost of modules (Fu et al., 2017)!





Literature Review

- **Vasseur and Kemp (2015)** surveyed 817 people in Netherlands that included 38 PV adopters and 779 non-adopters, and found that the cost of a PV system is the most important element. They then ran a logistic regression on attribution perception (with no controls), rather than attribution itself.
- **Sigrin et al. (2015)** surveyed 1234 adopters and 790 non-adopters in San Diego, and found descriptive differences between these two groups in terms of their income, education, and home size.
- **Bashiri and Alizadeh (2017)** surveyed 345 families (86 non-adopters) in Tehran and found seven important variables (family size, education, house ownership, etc.); however, income had a negative sign.
- **Fleiß et al. (2017)** surveyed 870 (potential) community solar participants in Austria and found that financial beliefs are the main driver for joining. Age effect was significant in the logit model. Results on gender, house ownership, education, and income were not consistent.
- Many other studies have focused on people's **motivations** (Islam and Meade, 2013; Balcombe et al., 2014; Schelly, 2014; Simpson and Clifton, 2017; Wolske et al., 2017), or adoption behaviors at a **regional** level (De Groote et al., 2016; Briguglio and Formosa, 2017).

Data

- Our research team surveyed 3,600 single-family households in four states (CA, AZ, NJ and NY) from June 2014 and April 2015 for both PV adoption and non-adoption status[1].
- Adopter samples primarily came from installer companies, and non-adopters from companies (i.e. lost leads) and an internet panel[2].

	California	Arizona	New Jersey	New York
Adopter	525	464	422	519
Non-adopter	1181	109	185	187

- After data cleaning for this research, we are left with 2549 samples.

1. Data from the study is publically available: https://data.nrel.gov/submissions/68
2. For further descriptive analysis see: Moezzi et al 2017 https://www.nrel.gov/docs/fy17osti/67727.pdf

Methods

- We compare two methods to predict solar adoption status
  - Logistic regression
  - Model-based machine learning (small sample not good for deep learning)
- We focus on ten highly visible and easy-to-measure attributes of correspondents in this binary classification task. Variables include electricity monthly bills, income, education, house sqft, family size, having kids or not, etc.
- The same variables are used for both two methods. 20% of sample is reserved for cross-validation.
- For ML, we employ the extreme gradient boosting model (XGBoost) that has won several recent Kaggle competitions.

Methods: XGBoost

- XGBoost is the extreme (fast) version of gradient boosting model (GBM), while GBM is an ensemble of trees, or more specifically a set of classification and regression trees (CART).

Model: assuming we have K trees

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}$$

Objective

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \qquad \Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

Training loss

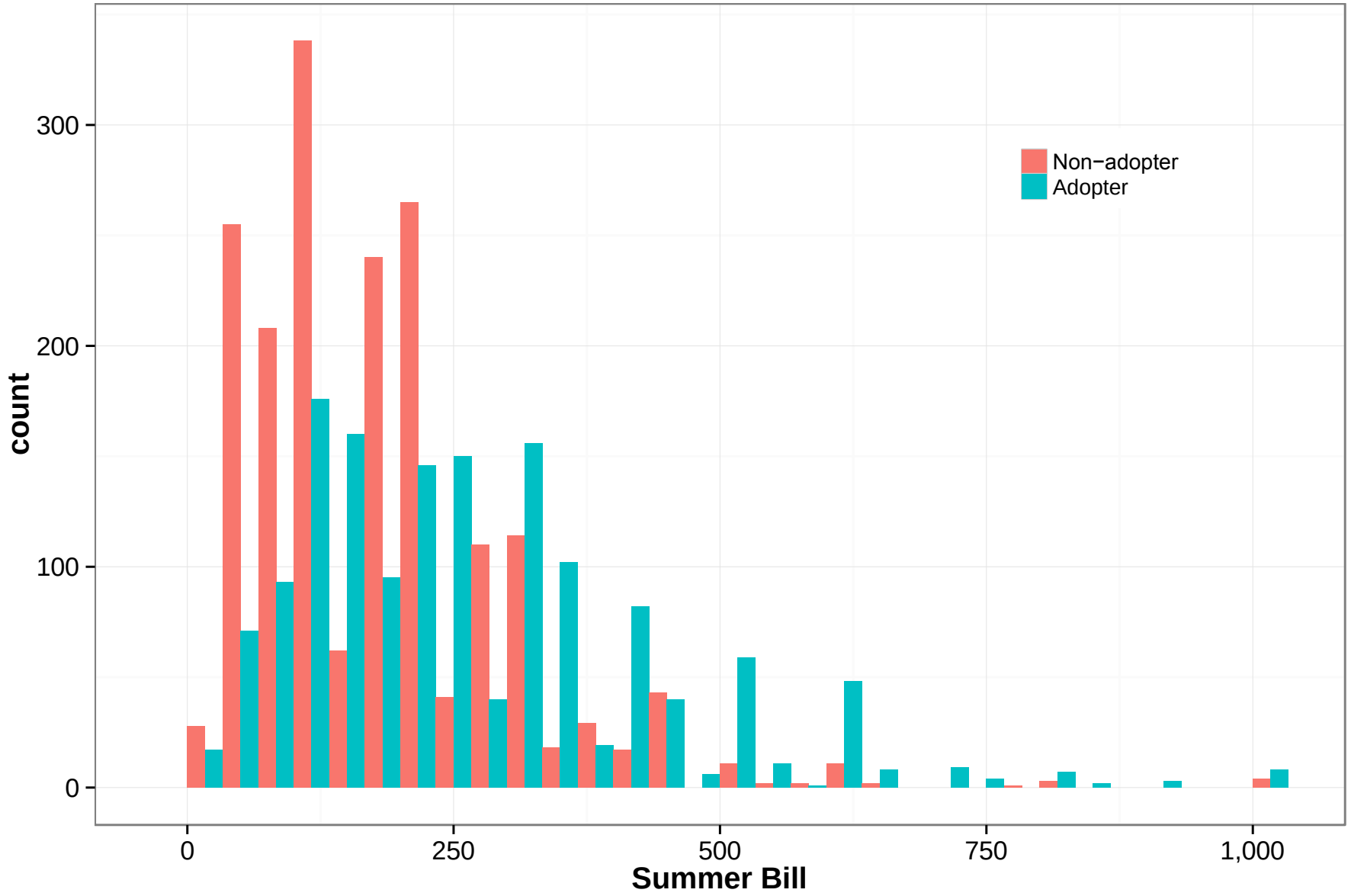
Complexity of the Trees

Number of leaves

L2 norm of leaf scores

Source: Chen (2014)

Results: Descriptive

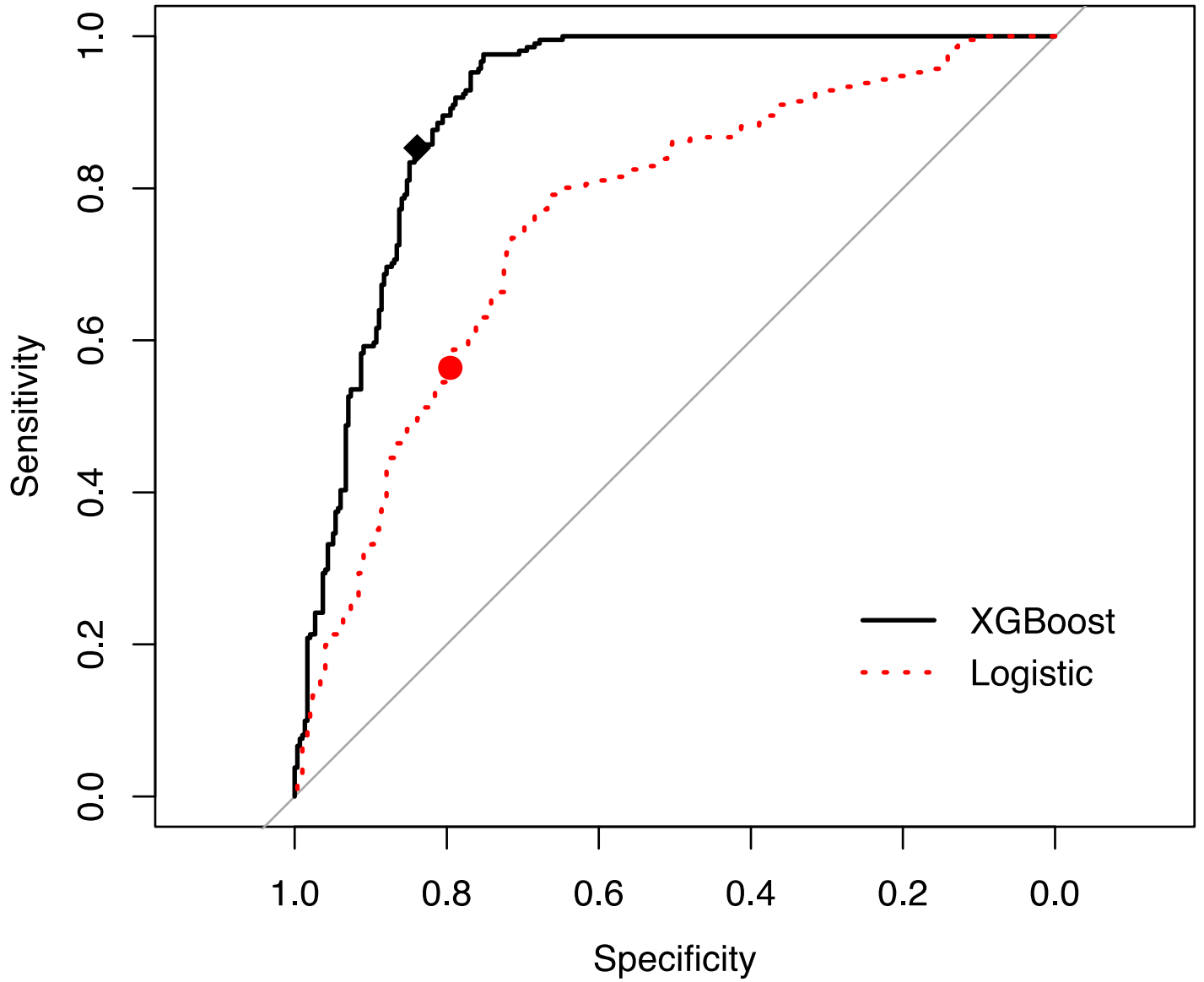


Results: Logistic Regression

Variable	Coefficient	Variable	Coefficient
Summer month bill	0.004***	Family size (#)	0.438***
Winter month bill	0.001	Having kids	0.425**
Income (in $1,000)	0.004***	Retire	-0.044
Some college	0.879***	Age	0.036***
Bachelor degree	1.166***	Independent	0.066
Graduate degree	0.997***	Republican	0.07
Sqft (1000, 1500)	0.914**	Sqft (3000, 4000)	-0.394
sqft (1500, 2000)	1.055***	Sqft (4000, 5000+)	-1.335*
sqft (2000, 3000)	0.591*		

Significance levels: *** 0.001; ** 0.01; * 0.05

Classification Results: ROC Curve Comparison



Sensitivity: $\Pr(\bar{Y} = 1 \mid Y = 1)$

Specificity: $\Pr(\bar{Y} = 0 \mid Y = 0)$

Sensitivity: 0.56 vs. 0.85

Specificity: 0.80 vs. 0.84

Conclusions

- We have shown and compared two methods to screen PV adopters and non-adopters in the U.S.: classical logistic regression and a machine learning method (XGBoost).
- ML is demonstrated to be significantly more powerful in classifying adoption likelihood, because ML is nonlinear and combines regression with regularization.
  - Deep learning is also feasible with more data inputs
- We intentionally exclude other variables identified in literature, e.g. perceptions and social values, as these “invisible” variables are not easily measured by solar companies.
  - Such variables might be measured indirectly e.g. purchased from data firms. Including them would likely increase our prediction capability.
- The PV industry could potentially make use of our ML method, since it can enhance the sensitivity (accuracy rate for adopters) by almost 30%.

Thank You

Questions?

Variable Importance by XGBoost