Predicitng Rooftop Solar Adoption

Donnie Meyer

November 10, 2018

Problem

- This project set out to build a predictive model for rooftop solar adoption
- There are many reasons why consumers choose to adopt rooftop solar
- Consumers adopt for economic benefits as well as environmental concern
- Knowing which factors are important can help solar companies target customers
- Better communication strategies can help push this important technology forward

Data Description

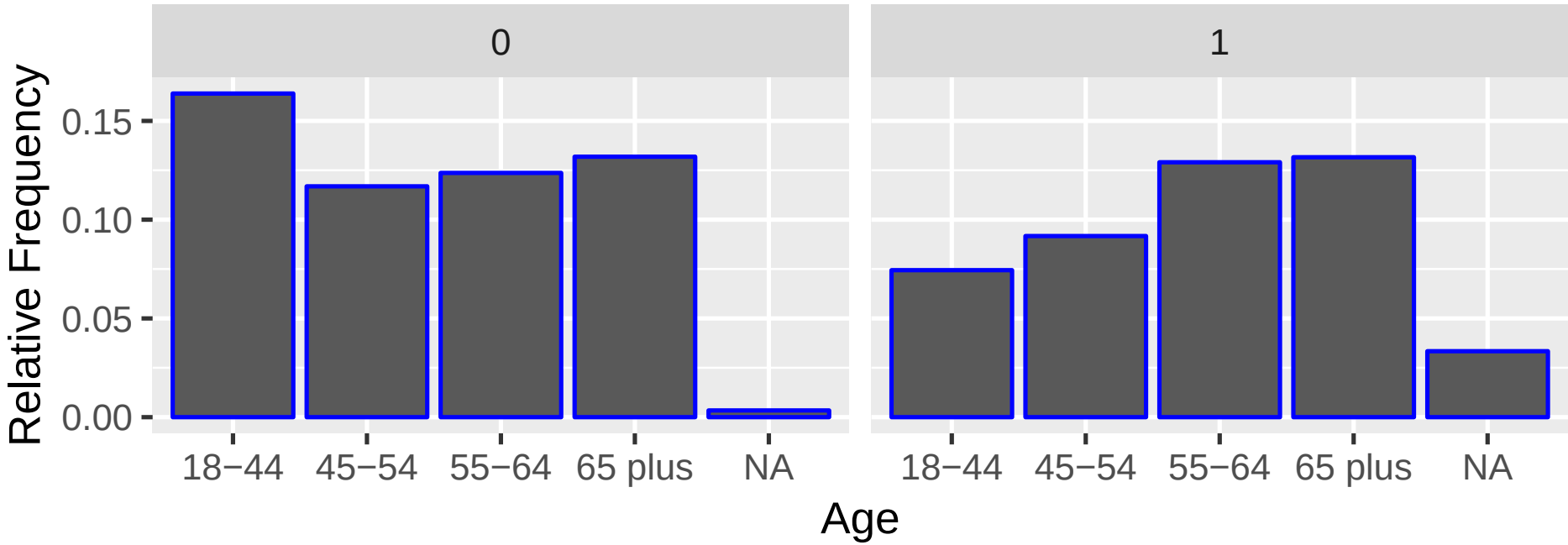
- Understanding the Evolution of Customer Motivations and Adoption Barriers in Residential Solar Markets: Survey Data
- Made available by the National Renewable Energy Laboratory (NREL)
- Survey data that captures demographics and customer behavior
- *HAVESOLAR* is a survey question asking "Do you currently have roof top solar installed at your current home"

Data Wrangling

- Appended three data sets
- Only those columns (survey questions) that where common across all three surveys were kept in final data set
- r-package *naniar* has a function called replace_with_na_all() that helped replace values of 99, 98, and 95 with NA
- Both unordered and ordered factor variables existed
- Numerous amounts of recoding and renaming was required

EDA Demographics

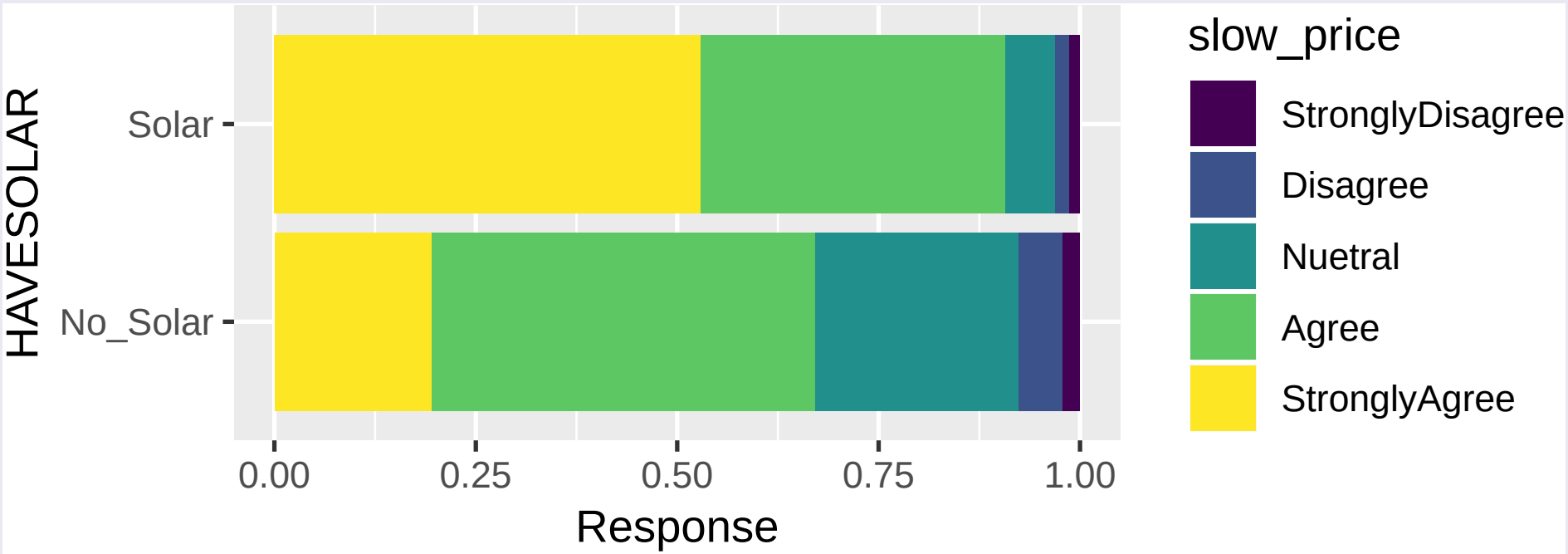
The first step of EDA was to compare the distributions between rooftop solar adopters and non-adopters. This is a comparison of the age distributions between the two groups. One observation is that only 28% of 18-44 year olds had adopted solar in this survey yet 50% of 65 plus had adopted. This suggests a relationship between age and adoption.



EDA Survey Responses

We can visualize survey questions in the form of stacked bar charts. Here we can see differences between solar adopters and non-adopters and their responses to survey questions. Those who strongly agree adopted solar much more than those who have not.

Using solar will help protect my family from rising electricity prices in the future



EDA Contingency Tables

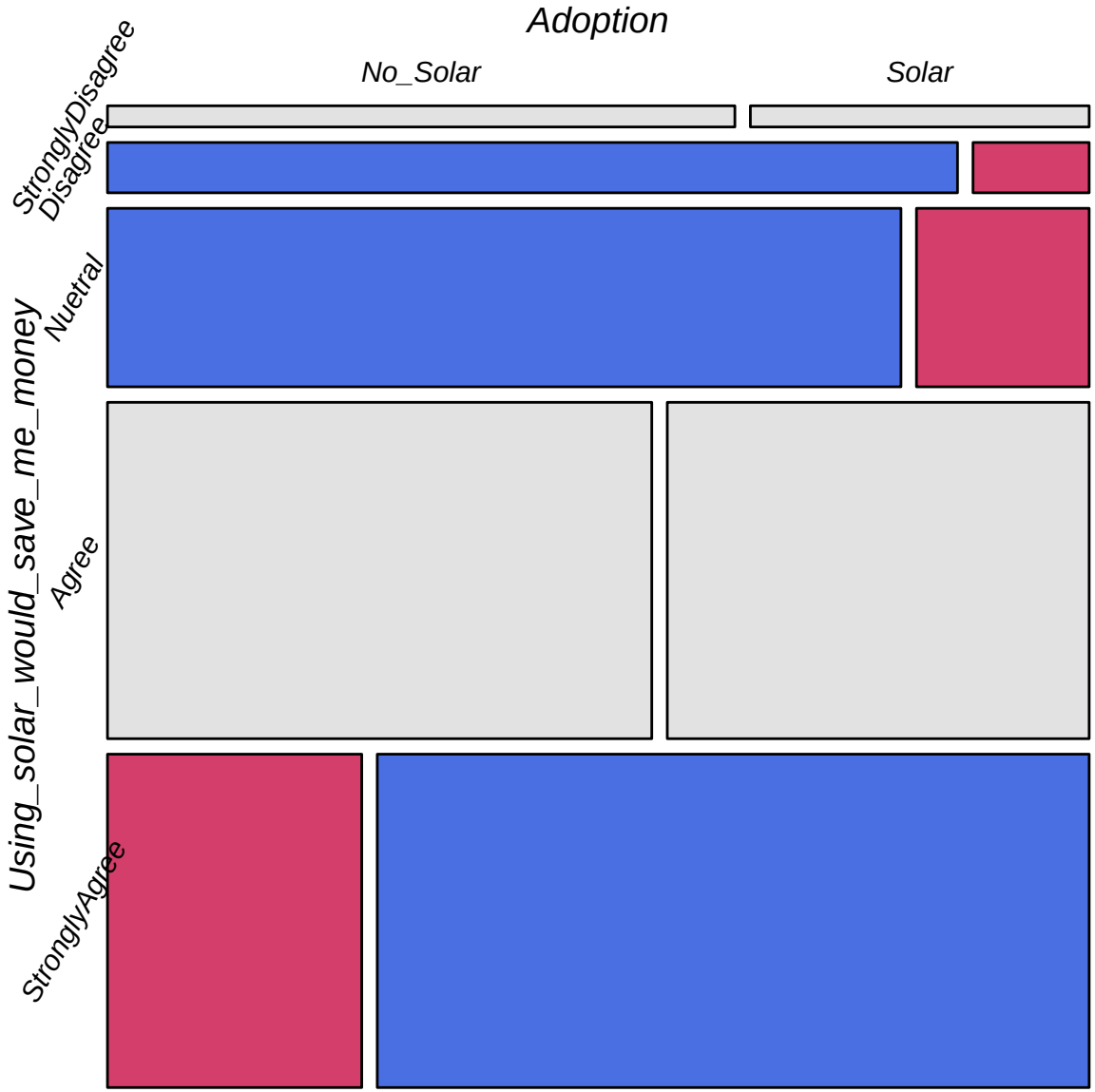
Contingency tables are a great way to view responses of survey questions between groups. They are also a necessary for conducting tests of independence such as the Pearson-Chi_Square Test.

Table 1: Using solar would save me money

	StronglyDisagree	Disagree	Nuetral	Agree	StronglyAgree
No_Solar	50	161	533	689	319
Solar	27	22	116	534	893

EDA Mosaic Plots

- Mosaic plots allow us to visualize contingency tables
- One way to think about mosaic plots is that they are scatter plots for categorical data
- The area of the square represents a conditional relative frequency
- Colored boxes represent Pearson residuals, indicating dependence between two groups
- Grey indicates independence between two groups
- This mosaic plot shows that there is a positive relationship between *HAVRSOLAR* and *save_money*


Adoption
No_Solar
Solar
Using_solar_would_save_me_money
StronglyDisagree
Disagree
Nuetral
Agree
StronglyAgree

EDA Independence Tests and Measures of Association

For nominal variables the Pearson's Chi-Square is used to test the independence between groups while Cramers V is used as the measure of Association. For ordinal variables we use the Cochran-Mantel-Haenszel test for testing independence and the GKgamma as the measure of association. Below we test the independence between *HAVESOLAR* and *save_money* as an example.

```
    Pearson's Chi-squared test

data:  save_money_tab
X-squared = 665.74, df = 4, p-value < 2.2e-16


                     X^2 df P(> X^2)
Likelihood Ratio 711.63  4        0
Pearson          665.74  4        0

Phi-Coefficient   : NA
Contingency Coeff.: 0.407
Cramer's V        : 0.446
```

Feature Selection

Below are handpicked features that all have statistical dependency with *HAVESOLAR* and are at least weakly associated.

Demographics

GENDER - GENDER, nominal
AGE_BINNED - age, ordinal
STATE - STATE, nominal
INCOME_BINNED, ordinal

Economic Variables

WINTER_NOPV_BINNED - winter_bill, ordinal
SUMMER_NOPV_BINNED - summer_bill, ordinal
BTE8 - slow_energy_price, ordinal
BE13 - return_investment, ordinal
*BE10 - save_money, ordinal, ordinal

Consumer Variables

CIJM1 - ask_someone_brand, ordinal
CIJM2 - ask_someone_service, ordinal
CIJM3 - trust_opinions, ordinal
CNS1 - look_new_products, ordinal
CNS2 - new_experience_products, ordinal
CNS3 - visit_places_products, ordinal

Environmental Variables

PN1 - renewable_energy, ordinal
BB1 - environment_improve, ordinal
BB2 - slow_climate_change, ordinal
BB3 - reduce_footprint, ordinal

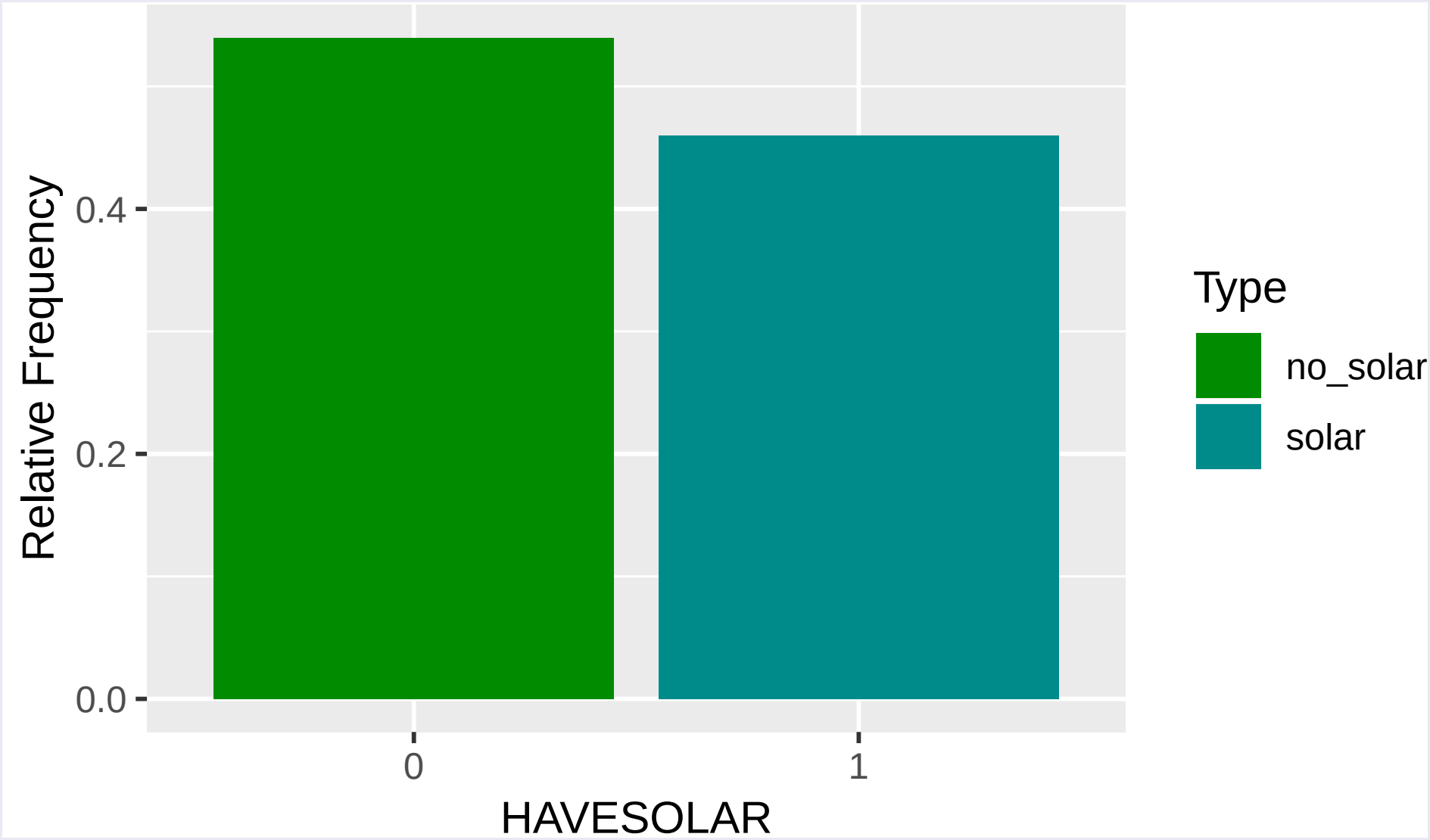
Baseline Method

For a classification problem, we predict the most frequent outcome for all observations. We see here that the base line model has an accuracy of 54 percent. A logistic regression model will be built in attempt to beat the base line prediction.

```
Adoption
No_Solar    Solar
    1907     1626
```

```
[1] 0.5397679
```

Dependent Variable



Training and Test Data

- The sample.split() function will be used which is part of the catools package in R
- The first argument is the dependent variable, and second argument is the percentage of data we want in the training set
- This also makes sure that the outcome variable is well balanced, as mentioned previously
- The training data will consist of 75% of the original data with the remaining 25% saved for the test data

-training data -test data

Logistic Regression

- Logistic regression predicts the probability of an outcome variable being true
- The logistic regression model will predict the probability that the consumer has adopted solar i.e. P(y=1)
- P(y=0) = 1 - P(y=1) where the P(y=0) is the probability the consumer has not adopted solar
- Positive parameter estimates are predictive of class 1, while negative values are predictive of class 0

Multicolinearity

Table 2: VIF Values

	GVIF	Df	GVIF^(1/(2*Df))
ordered(INCOME_BINNED)	1.584449	5	1.047099
factor(GENDER)	1.164484	2	1.038803
ordered(age)	1.438599	3	1.062486
factor(STATE)	1.508333	3	1.070901
ordered(winter_bill)	4.266467	10	1.075235
ordered(summer_bill)	4.541089	10	1.078594
ordered(slow_energy_price)	7.273596	4	1.281500
ordered(return_investment)	5.663080	4	1.242029
ordered(save_money)	8.066370	4	1.298180
ordered(ask_someone_brand)	8.757583	4	1.311590
ordered(ask_someone_service)	10.777963	4	1.346068
ordered(trust_opinions)	5.499642	4	1.237490
ordered(look_new_products)	11.008798	4	1.349639
ordered(new_experience_products)	12.326236	4	1.368844
ordered(visit_places_products)	5.817539	4	1.246213
ordered(renewable_energy)	5.603086	4	1.240376
ordered(slow_climate_change)	4.488738	4	1.206467
ordered(reduce_footprint)	6.991003	4	1.275168

Wald Tests

We can test for the overall significance for each categorical variable using the Wald Test. In the table below we see that the only variable that is insignificant is the *slow_climate_change* variable.

	Df	Chisq	Pr(>Chisq)
ordered(INCOME_BINNED)	5	31.543480	0.0000073
factor(GENDER)	2	29.035433	0.0000005
ordered(age)	3	21.734946	0.0000741
factor(STATE)	3	201.358561	0.0000000
ordered(winter_bill)	10	40.166457	0.0000158
ordered(save_money)	4	172.655883	0.0000000
ordered(ask_someone_service)	4	45.216516	0.0000000
ordered(look_new_products)	4	48.050844	0.0000000
ordered(renewable_energy)	4	23.537781	0.0000989
ordered(slow_climate_change)	4	7.047388	0.1334051

Predicted Values

Lets see if we are predicting higher probability for HAVESOLAR = 1, and lower probability for HAVESOLAR = 0. We are predicting an average probability of 0.74 for HAVESOLAR = 1 and 0.20 for HAVESOLAR = 0. This a good sign because we are predicting a higher probability for the actual HAVESOLAR = 1 cases.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.09378 0.38159 0.44168 0.80174 0.99321
```

```
        0         1
0.2195201 0.7225034
```

Confusion Matrix

- TN cases predict $HAVESOLAR = 1$ and the case is $HAVESOLAR = 0$
- TP cases predict $HAVESOLAR = 1$ and the case is $HAVESOLAR = 1$
- FN cases predict $HAVESOLAR = 0$ and the case is $HAVESOLAR = 0$
- FP cases predict $HAVESOLAR = 0$ and the case is $HAVESOLAR = 1$

Table 3: Predict HAVESOLAR, t = 0.5

	Predicted = 0	Predicted = 1
Actual = 0	908	145
Actual = 1	165	668

Measures of Performance

From the confusion matrix we can calculate various measures. I will focus on sensitivity, specificity, and over all accuracy.

sensitivity - the percentage of TRUE cases classified correctly

```
[1] 0.8019208
```

specificity - the percentage of FALSE cases classified correctly

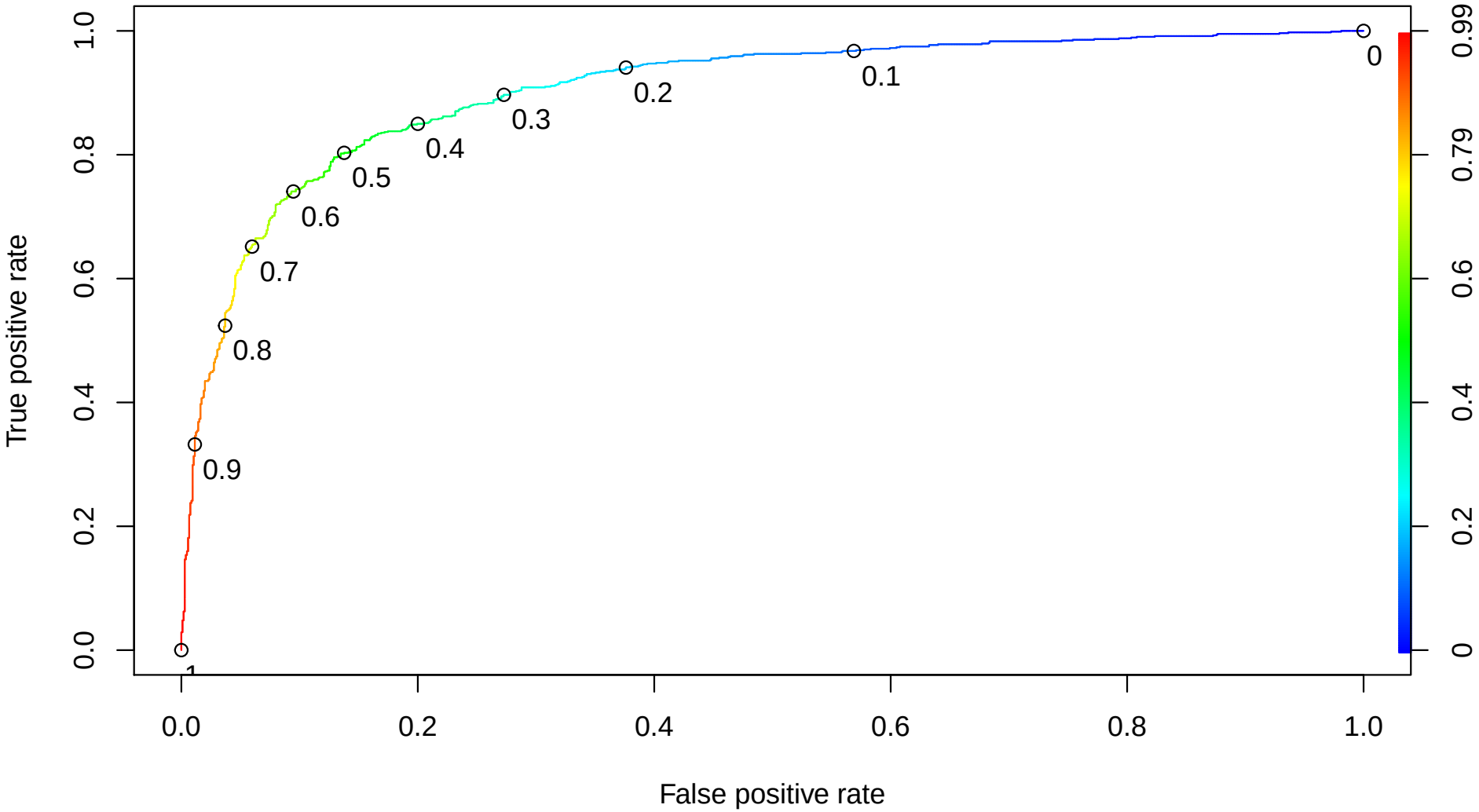
```
[1] 0.8622982
```

overall accuracy - the percentage of overall cases classified correctly

```
[1] 0.835631
```

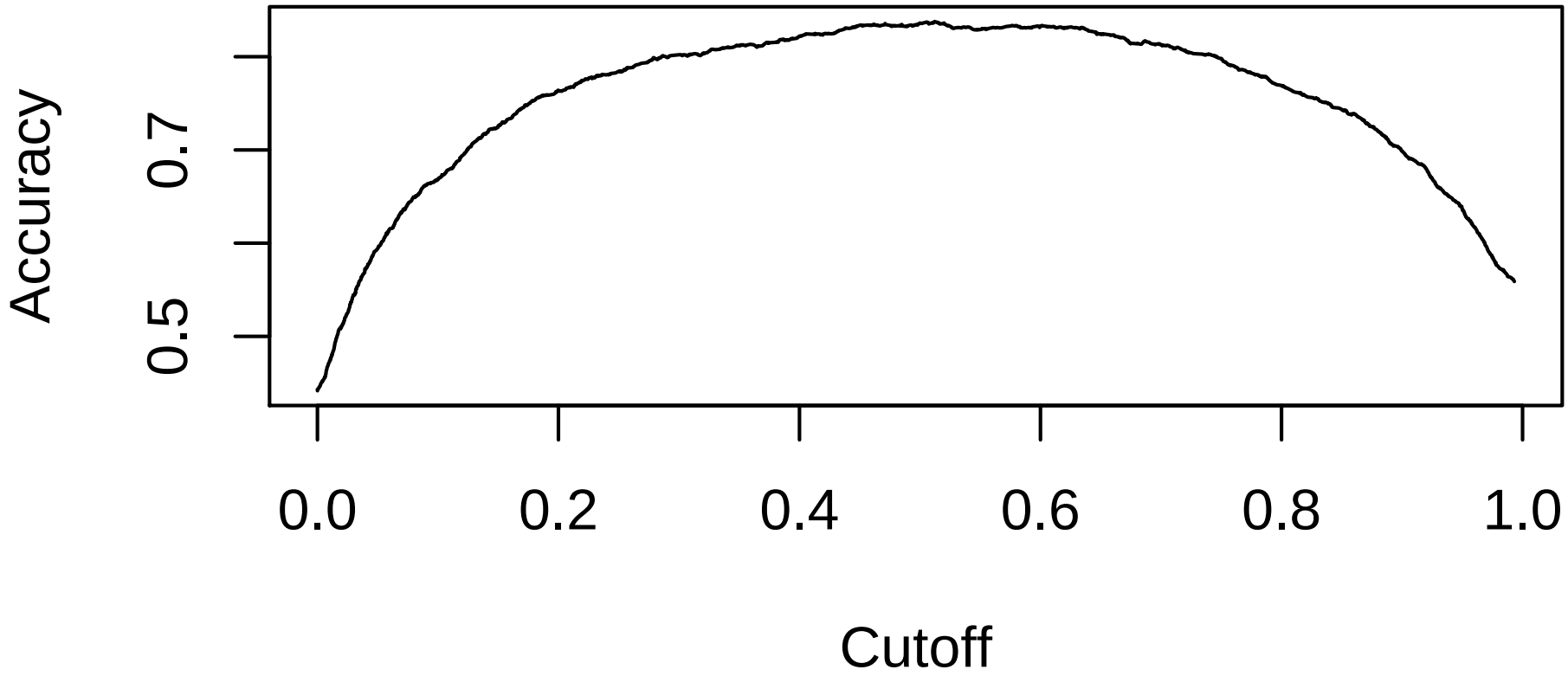
ROC

- The ROC curve has a true-positive rate on the y-axis and a false positive rate on the x-axis
- $\mathbf{t} = 1$ is at the point (0,0) while $\mathbf{t} = 0$ is at the point (1,1)
- The ROC curve is a performance measure of a classification model at all thresholds $\mathbf{t}$
- We will use the ROC to calculate the AUC value


True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.99
0.79
0.6
0.4
0.2
0
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

Calculating t for highest accuracy

R allows us to calculate the accuracy, the cutoff, and the ability to visualize this. Below is a plot and cutoff that maximizes accuracy to determine our threshold value **t**.




```
##     accuracy cutoff.1228
##    0.8377519   0.5121446
```

AUC

- The AUC provides an aggregate measure of performance across all possible classification thresholds
- It measures the probability that a random TRUE value, in our case $HAVESOLAR = 1$, is to the right of a random NEGATIVE value $HAVESOLAR = 0$
- 90% of the time our random true value is to the right of our random negative value
- A simpler interpretation is that when the AUC is 0 we have predicted none of our values, and when the AUC is 1 we predict 100 % of our values correctly.

```
Area under the curve: 0.9043
```

Test Set Precition Accuracy

The measures are very close indicating our model is doing a good job of predicting *HAVESOLAR* in the test data.

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0000079 0.0830335 0.3483680 0.4355143 0.8089502 0.9957319
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.09378 0.38159 0.44168 0.80174 0.99321
```

Test Set Confusion Matrix

The model is performing well on the test data with an overall accuracy of 82%.

Table 4: Predict HAVESOLAR, t = 0.5

	Predicted = 0	Predicted = 1
Actual = 0	297	54
Actual = 1	60	218

Test Set Performance Measures

sensitivity - the percentage of TRUE cases classified correctly

```
[1] 0.8461538
```

specificity - the percentage of FALSE cases classified correctly

```
[1] 0.7841727
```

overall accuracy - the percentage of overall cases classified correctly

```
[1] 0.8187599
```

Test Set Accuracy

We have an AUC of 89% so our model seems to be performing well. The model is predicting 89% of the $HAVESOLAR = 1$ cases correctly.

```
Area under the curve: 0.8938
```

Reccomendations

Economic Variables

- It is apparent that economic factors play the strongest role in rooftop solar adoption
- Adopters believe that rooftop solar saves them money and protects them from rising energy prices in the future
- Solar companies should continue to educate potential customers about the economic benefits of adopting rooftop solar - This can be done through advertising, social media, information on websites etc.
- Solar companies should target areas where electricity bills are high and solar generation potential is possible
- Consumers find it easy to understand savings when using their energy bills
- Information should be made available about how consumers will lower monthly utility bills when adopting solar

Consumer Variables

- Predictors from the consumer category **included ask_some_service**, **look_new_products**, and **visit_places_products** are negatively correlated with *HAVESOLAR*
- This implies that solar adopters are not actively looking or seeking new products, do not inquire with others about new products, and do not go to new places seeking new products
- This means that most potential solar adopter may be harder to reach through traditional advertising strategies
- Efforts to reach these customers could be essential and more research should be done to understand this behavior

Enviromental Variables

- Environmental reasons also play a factor
- The most highly correlated factor from the environmental category was **renewable_energy**, though factors where correlated with *HAVESOLAR**
- Solar companies should maintain this image of being green
- May encourage those who already know of the financial benefits to ultimately make the switch

Demographic Variables

- Men adopted solar more than females
- California adopted solar more than the other states
- Older aged consumers adopted solar more than younger consumers
- Consumers with higher incomes adopted more solar than lower income consumers
- When developing marketing strategies keep these demographic factors in mind.