Date: 27th-01-2021

Author : Bessam Mehenni

Adoption of solar energy in residential in Virginia state

1. References

[1] Data source: Stanford's Deepsolar dataset, dec.2018

http://web.stanford.edu/group/deepsolar/home

[2] Python notebook / data analysis:

https://www.kaggle.com/andromedasagan/implementation-of-solar-energy-in-the-us

2. Context

The source of data is Stanford University's DeepSolar project, a deep learning framework that analyzed satellite images to detect solar panels throughout the country. The data collected are the size, type (residential/non-residential) of the power systems distributed in the 48 states in the U.S. The associated socioeconomic data for these locations were recorded over several years.

My ambition in this work is to build a socio-economic analysis of the last mile to understand what are the profiles within a homogeneous group of households that is adopting solar energy. It focuses on the state of Virginia.

This work is based on a first chapter of data analysis [2] that highlights key trends and correlations in the deployment of solar power based on the full Deepsolar dataset.

3. Goal

- Explore the data, handle missing values.
- Make visualizations to identify trends. Identify the characteristics of a homogeneous group of households that makes the majority of installed systems.
- Model by ML the adoption of solar energy by households. we take the target variable **Solar_panel_area_per_capita** to illustrate the adoption of solar systems. Using ML to show the factors involved in explaining the adoption.

#Explanations about the way I proceed:
#Importing Deepsolar dataset restricted to Virginia state
#Creation of column 'employment_rate' and calculation of the employment rate
#Deleting 'employed' and 'unemployed' columns

	tile_count	solar_system_count	total_panel_area	fips	average_household_income	county	education_bachelor
7529	9.0	4.0	1361.899130	51710980300	NaN	Norfolk city	0
51696	17.0	15.0	1059.199094	51107980100	NaN	Loudoun County	0
12059	1.0	1.0	17.092393	51059980100	NaN	Fairfax County	0
49028	26.0	8.0	2180.313599	51087980100	NaN	Henrico County	0

4 rows × 168 columns

4. Data exploration

We will measure the **adoption of solar systems** through the variable

`solar_panel_area_per_capita`. We will draw up a matrix of correlations of deemed and less deemed factors over the target.

Certain factors are deemed to be decisive in the choice for households to equip with solar systems. These factors become evident when the following observations are made:

- households are likely to equip themselves with solar equipment where the solar resource is the most abundant.
- households that can afford it financially are more likely to invest in solar systems.
- incentives can facilitate access to solar systems, especially for households that initially lacked the capacity to afford them.
- feed-in tariffs for grid electricity are decisive for acquiring solar systems.

Two missing factors that will unfortunately not be studied here, as they have not been collected in the present dataset:

- I think that data about the **intensity of incentives** would have been useful, i.e. special retail electricity tariffs (in c$/kWh) or investment incentives for the purchase of solar systems. This can make the difference and allow households to move from a situation of non-accessor to potential accessor of a solar system.
- the **ecological awareness** is a more personal factor that is likely to be influential. Maybe it could have been read in voting intentions for example.

#Explanations about the way I proceed:
#Variables are renamed : 'incentive_count_residential' in 'incentive_count_resid', 'incentive_residential_state_level' in 'incentive_count_resid_state'

Short description of the extracted dataset

Dataset contains socio-economic and environmental data.

- county - county name
- average_household_income - average annual household income ($),ACS 2015 (5-Year Estimates)

- daily_solar_radiation - daily solar radiation (kWh/m^2/d),NASA Surface Meteorology and Solar Energy
- incentive_count_resid - number of incentives for residential solar,www.dsireusa.org
- avg_electricity_retail_rate - average residential retail electricity price over the past 5 years,EIA
- incentive_count_resid_state - number of state-level incentives for residential solar,www.dsireusa.org
- solar_panel_area_per_capita - solar panel area per capita (m^2/capita),deepsolar
- age_median - median age,ACS 2015 (5-Year Estimates)
- number_of_years_of_education - number of years of education,ACS 2015 (5-Year Estimates)
- employed,number of employed people,ACS 2015 (5-Year Estimates)
- unemployed,number of unemployed people,ACS 2015 (5-Year Estimates)

5. Missing values

In the dataset, there is a problem with very large floating-point numbers for which INF values are returned. To solve the problem, the max values are filtered and then discarded.

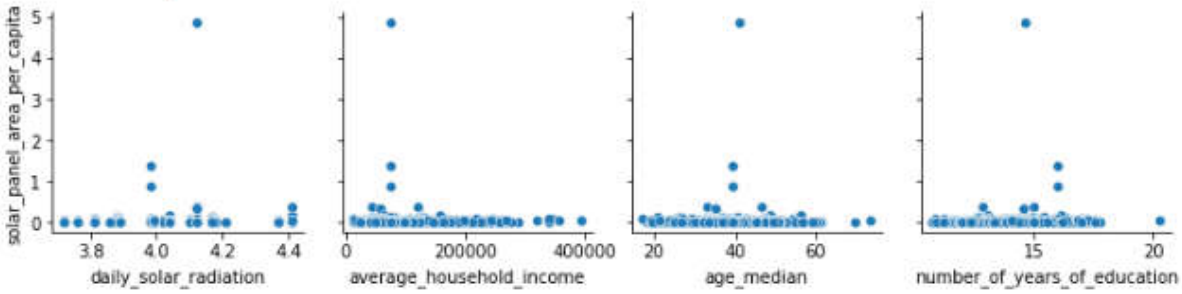
#Explanations about the way I proceed:
#Data cleaning: deletion of values Inf
#Printing a dataset describe

	average_household_income	daily_solar_radiation	solar_panel_area_per_capita	age_median	number_of_years_of_education	incentive_count_resid
count	1870.000000	1709.000000	1879.000000	1877.000000	1879.000000	1902.0
mean	88862.323262	4.036934	0.013861	39.022536	13.918186	19.0
std	46959.870396	0.117674	0.120201	7.045672	1.408296	0.0
min	13482.837838	3.720000	0.000000	17.200000	10.770450	19.0
25%	55114.086161	3.980000	0.000000	34.000000	12.801099	19.0
50%	75336.806575	3.990000	0.003541	39.300000	13.739533	19.0
75%	111138.929886	4.120000	0.012778	44.000000	14.921504	19.0
max	394499.587203	4.410000	4.877326	73.500000	20.263158	19.0

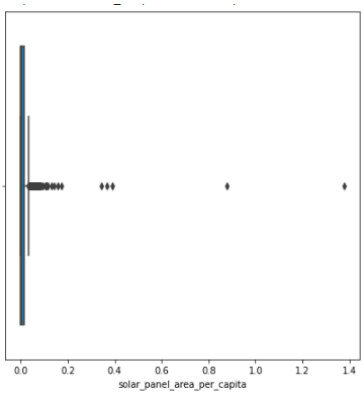
#Fill NaN with the column median value (except in`daily_solar_radiation` with the mean value)
#Delete rows without values in `Solar_panel_area_per_capita`.

6. Vizualisations

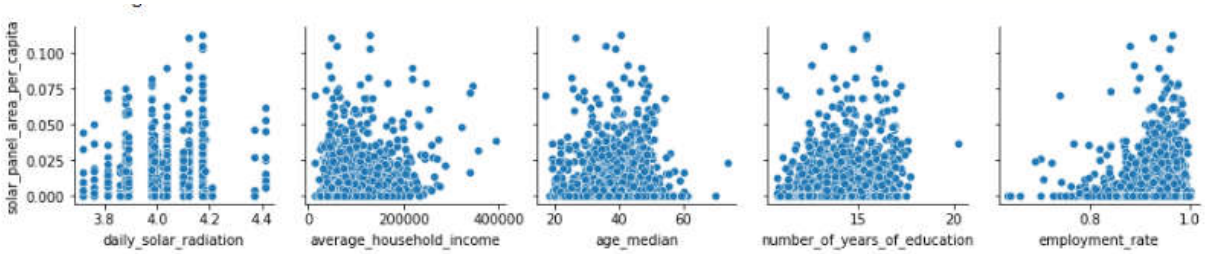
#Visualization of the different behaviors between the main factors and the target in order to identify po ssible outliers _



#Data cleaning: deletion of outliers



#Data cleaning: We see several outliers that can be deleted after 0.125



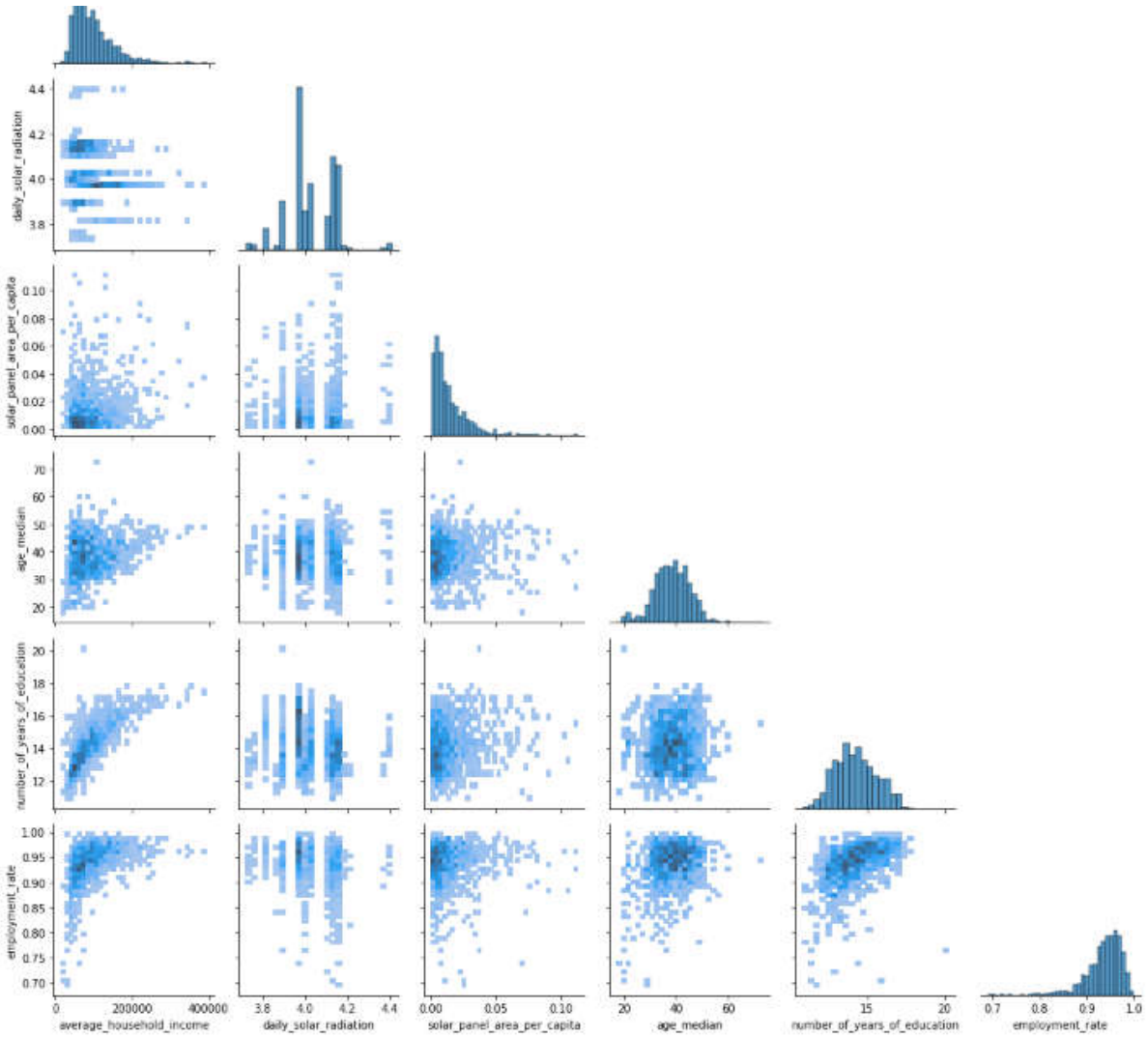
We take the target variable `Solar_panel_area_per_capita` to illustrate the **adoption of solar systems**.

#To not skew the analysis by zero-values, we are keeping a dataset with only rows `'solar_panel_area_per_capita'` > 0

#At this time, we ignore variables that are identical in value within the entire population (std = 0)

#Drop_elements = `['county','incentive_count_resid','incentive_count_resid_state','avg_electricity_retail_rate']`

Putting the factors face to face on graphs will allow us to see how what the locality has in household profiles contributes to adoption.



According to the `Solar_panel_area_per_capita` histogram, the group of vast majority of installed solar systems is located where `Solar_panel_area_per_capita` is **below 0.03 m²/capita**.

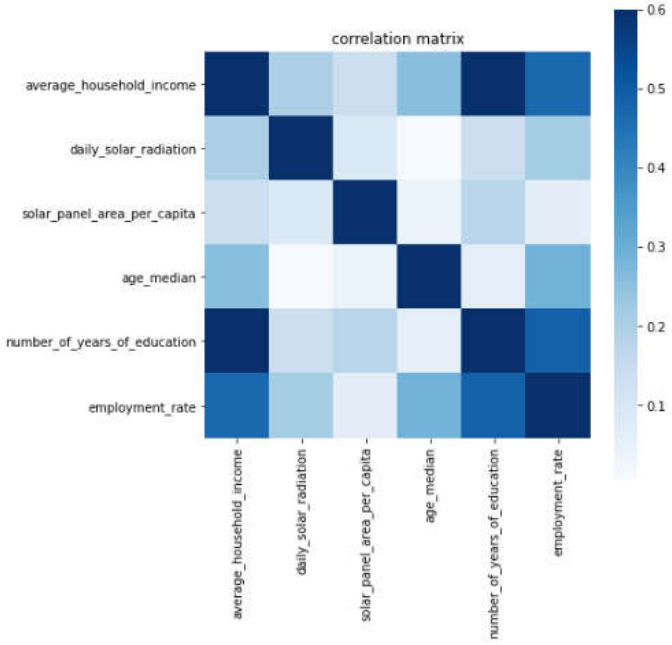
	average_household_income	daily_solar_radiation	solar_panel_area_per_capita	age_median	number_of_years_of_education	employment_rate
count	954.000000	954.000000	954.000000	954.000000	954.000000	954.000000
mean	95328.249694	4.029536	0.010629	38.472222	14.170652	0.934923
std	47320.275286	0.110296	0.007482	6.687310	1.330418	0.040264
min	15303.669725	3.720000	0.000453	19.100000	10.770450	0.689376
20%	56042.616811	3.980000	0.003896	33.060000	12.962141	0.909664
50%	83617.439560	4.000000	0.008675	38.500000	14.060450	0.941719
80%	128031.939049	4.120000	0.017255	44.000000	15.396677	0.966375
max	340033.097970	4.410000	0.029979	73.500000	17.743657	1.000000

Let's look at the ranges of median age, income and education level of 60% of the records of this sample, between the 20% and 80% percentiles.

We note the very meaningful characteristics of the group of households whose `Solar_panel_area_per_capita` is below 0.03 m^2/capita for the vast majority of installed solar systems. For **60% of the records in this sample**:

- average household income is between **56000** and **128000**$
- median age is between **33** and **44 years old**.
- level of education is between **13** and **16 years of education**.

7. Correlational analysis



There is a very strong correlation between `average_household_income` and `number_of_years_of_education`. We can easily explain this correlation. In general, education opens doors to higher-paying jobs.

`Age_median` is much less correlated with `average_household_income`.

It would have been interesting to see how these factors correlate with other factors that reflect the "green" mentality or the motivation to do savings.

#What are the features without variance?

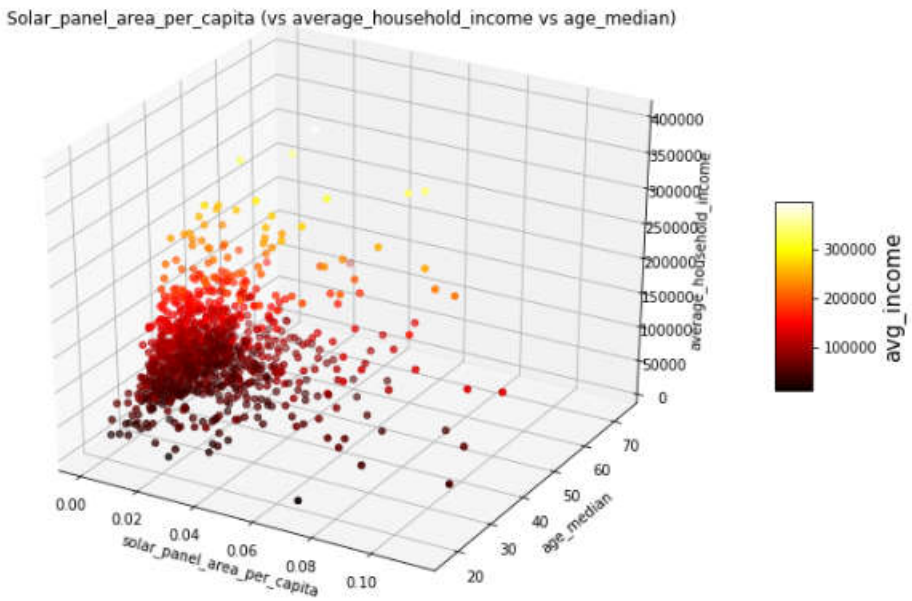
	average_household_income	daily_solar_radiation	solar_panel_area_per_capita	age_median	number_of_years_of_education	employment_rate
count	1102.000000	1102.000000	1102.000000	1102.000000	1102.000000	1102.000000
mean	97185.391151	4.032548	0.015619	38.516334	14.244628	0.935740
std	51285.993888	0.112948	0.015891	6.843867	1.370655	0.040540
min	14078.220140	3.720000	0.000453	17.200000	10.770450	0.689376
25%	60488.212602	3.980000	0.005142	33.900000	13.220098	0.918447
50%	84660.613082	4.020000	0.010245	38.650000	14.135445	0.943064
75%	119495.190419	4.120000	0.020506	43.300000	15.235297	0.964033
max	394499.587203	4.410000	0.112775	73.500000	20.263158	1.000000

#We ignore variables without variance that will not bring to a modeling.
#Despite there is a strong correlation of the target with `number_of_years_of_education`, the model gives better prediction results with it, that's why we keep it. And it works better without `'employment_rate'`.
#retained = `['number_of_years_of_education','daily_solar_radiation', 'age_median','average_household_income']`

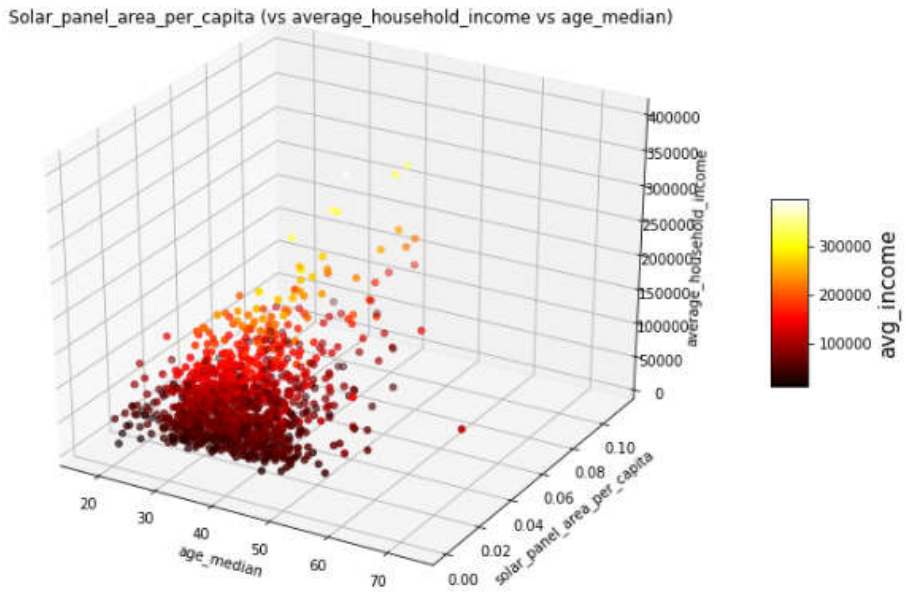
8. Forecasting solar adoption

8.1 Preliminary modeling

#3D visualization between selected factors in a first view



#3D visualization between selected factors in a second view



#Preliminary modeling using a RandomForestRegressor

```
Train score: 0.6196504573535415
Test score: 0.11404366275527165
```

Good fitting on the training data. The test score could be much higher.

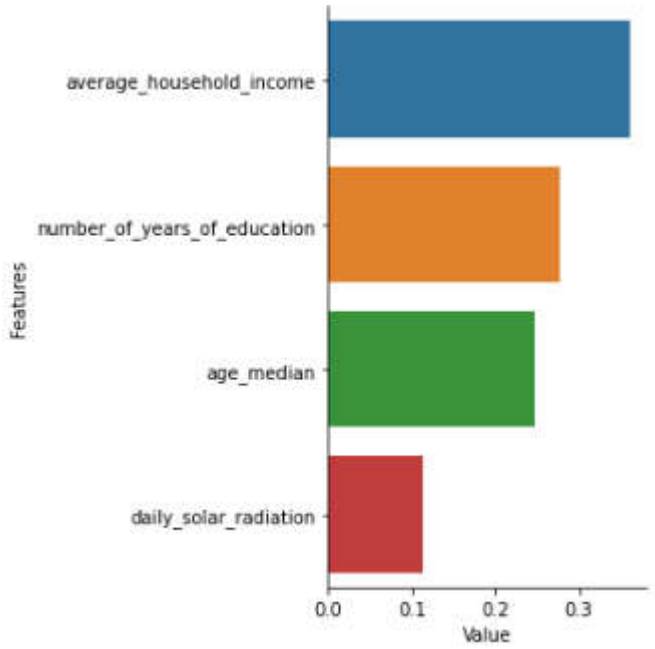
#Feature importance: ranking of the features in the explanation of the target

	Features	Value
3	average_household_income	0.361677
0	number_of_years_of_education	0.277283
2	age_median	0.247928
1	daily_solar_radiation	0.113112

A limited number of descriptives variables can acceptably explain the adoption of solar energy: the predominant are the `average_household_income` of the locality, `age_median` and `number_of_years_of_education` factors.

The level of education (`number_of_years_of_education`) factor may be influencing the adoption because of partly an intrinsec factor which is the level of income again (`average_household_income`). We have seen that these factors are highly correlated.

`Age_median` has also a contribution to the adoption.



#We continue with the same descriptive variables for the fine-tuned model
#retained = ['average_household_income','age_median','daily_solar_radiation','number_of_years_of_education']

8.2 Fine-tuning the parameters of the model

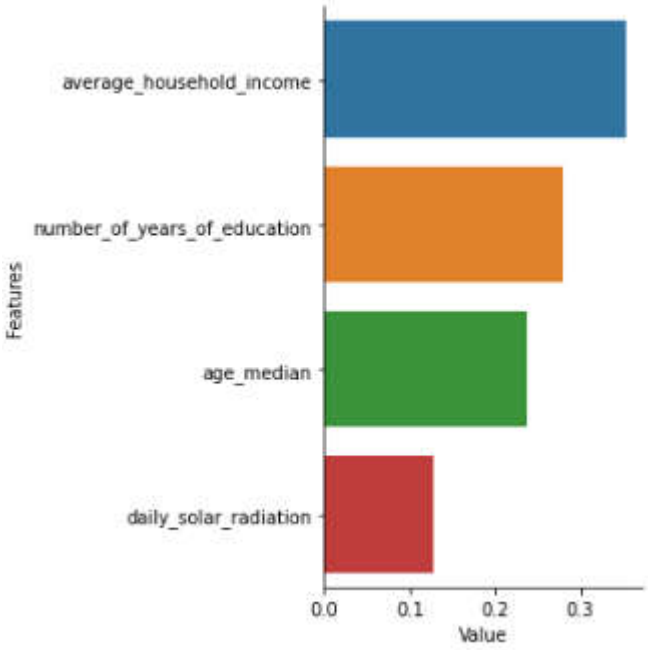
#GridSearch

```
Gridsearch train score: 0.5187964402125087
Gridsearch test score: 0.1494869909937131
```

Correct fitting on the training data, the test score is acceptable. The model shows some signs of overfitting, i.e. weaknesses in its generalizability.

#Feature importance : ranking of the features in the explanation of the target

	Features	Value
0	average_household_income	0.354072
3	number_of_years_of_education	0.278941
1	age_median	0.238171
2	daily_solar_radiation	0.128816



8.3 Conclusions

The fine-tuned model confirms that the `average_household_income` of the locality, `number_of_years_of_education`, `age_median` and `daily_solar_radiation` factors can acceptably explain the adoption of solar energy in Virginia state.

The `average_household_income` factor has the most important contribution.

The level of education (`number_of_years_of_education`) is also an important factor in which we can find intrinsically the level of income (`average_household_income`), as seen before in the correlation matrix.

Concerning the contribution of `age_median`, to explain this we should rather look at what advancement in age and career brings: perhaps the rationality of the choices of household members, the stability of its income and the ability to make investments in order to project savings in the upcoming years.