# CLASSIFICATION MODEL TO PREDICT THE PURCHASE OF SOLAR PANEL BY THE CUSTOMER

**By: SUKRITI MACKER (sukriti.macker98@gmail.com)**

- **In the following project the objective is to analyze and evaluate the data provided to us, to predict whether or not a customer would invest in solar panels.**
- **Also, the objective is to find out which state(s) of India would be appropriate and yield maximum profit, if chosen, to expand company's business.**

## The major steps to analyze and predict solutions from the data:-

1. IMPORTING LIBRARIES AND DATA
2. MANIPULATING THE DATA
3. BALANCING THE IMBALANCED DATA
4. SELECTING THE BEST FEATURES
5. DATA VISUALIZATION
6. DATA PREDICTION AND MODEL CREATION
7. CONCLUSION

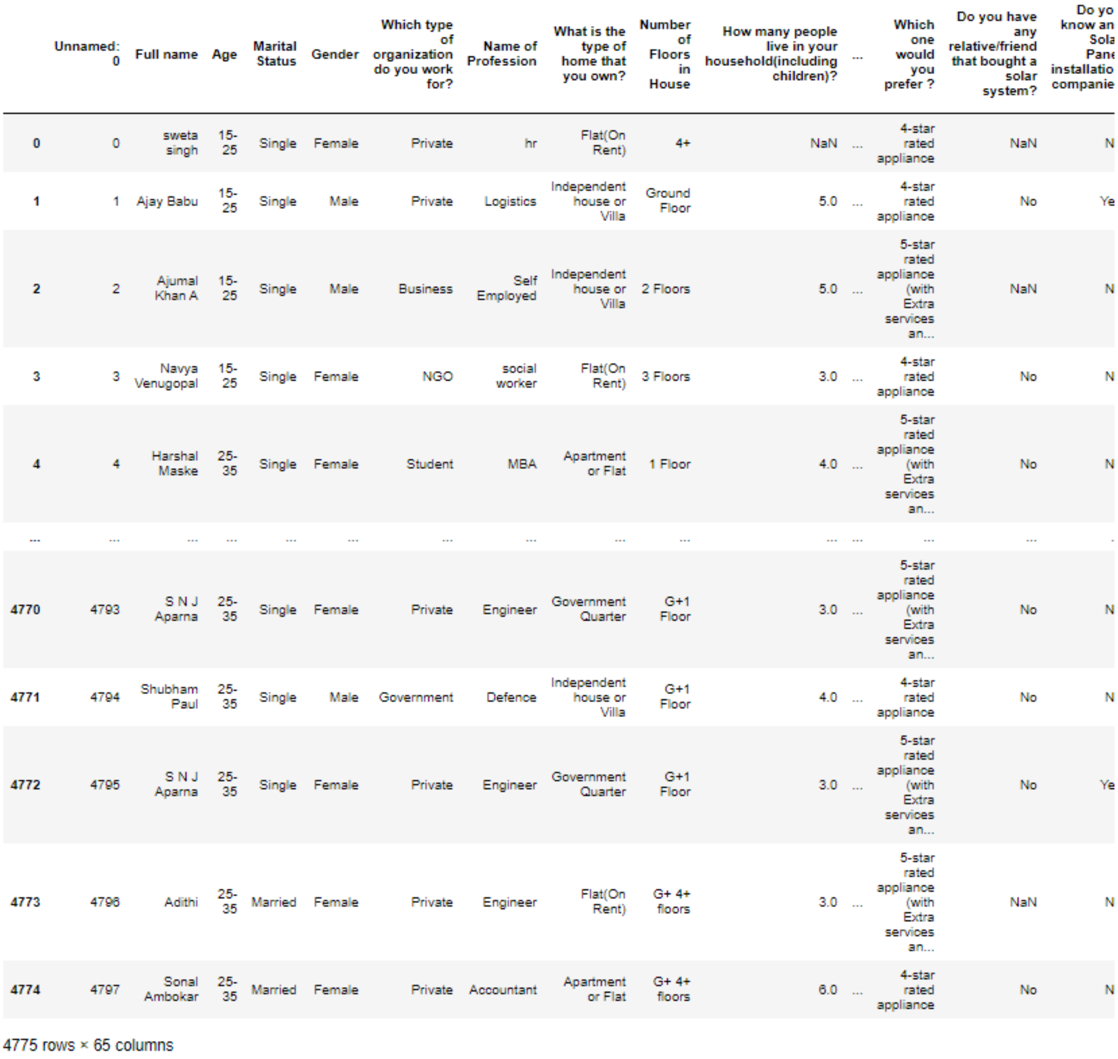
## 1. IMPORTING LIBRARIES AND DATA

We will import the pandas library, matplolib library and numpy library as follows:-

```
import pandas as pd

import matplotlib.pyplot as plt

import numpy as np
```

Now we will import the data file:-

```
data=pd.read_excel("Data Analytics Work.xlsx")
```



| | Unnamed: 0 | Full name | Age | Marital Status | Gender | Which type of organization do you work for? | Name of Profession | What is the type of home that you own? | Number of Floors in House | How many people live in your household(including children)? | ... | Which one would you prefer ? | Do you have any relative/friend that bought a solar system? | Do yo know an Sola Pane installatio companie |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | sweta singh | 15-25 | Single | Female | Private | hr | Flat(On Rent) | 4+ | NaN | ... | 4-star rated appliance | NaN | N |
| **1** | 1 | Ajay Babu | 15-25 | Single | Male | Private | Logistics | Independent house or Villa | Ground Floor | 5.0 | ... | 4-star rated appliance | No | Ye |
| **2** | 2 | Ajumal Khan A | 15-25 | Single | Male | Business | Self Employed | Independent house or Villa | 2 Floors | 5.0 | ... | 5-star rated appliance (with Extra services an... | NaN | N |
| **3** | 3 | Navya Venugopal | 15-25 | Single | Female | NGO | social worker | Flat(On Rent) | 3 Floors | 3.0 | ... | 4-star rated appliance | No | N |
| **4** | 4 | Harshal Maske | 25-35 | Single | Female | Student | MBA | Apartment or Flat | 1 Floor | 4.0 | ... | 5-star rated appliance (with Extra services an... | No | N |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | . |
| **4770** | 4793 | S N J Aparna | 25-35 | Single | Female | Private | Engineer | Government Quarter | G+1 Floor | 3.0 | ... | 5-star rated appliance (with Extra services an... | No | N |
| **4771** | 4794 | Shubham Paul | 25-35 | Single | Male | Government | Defence | Independent house or Villa | G+1 Floor | 4.0 | ... | 4-star rated appliance | No | N |
| **4772** | 4795 | S N J Aparna | 25-35 | Single | Female | Private | Engineer | Government Quarter | G+1 Floor | 3.0 | ... | 5-star rated appliance (with Extra services an... | No | Ye |
| **4773** | 4796 | Adithi | 25-35 | Married | Female | Private | Engineer | Flat(On Rent) | G+ 4+ floors | 3.0 | ... | 5-star rated appliance (with Extra services an... | NaN | N |
| **4774** | 4797 | Sonal Ambokar | 25-35 | Married | Female | Private | Accountant | Apartment or Flat | G+ 4+ floors | 6.0 | ... | 4-star rated appliance | No | N |

4775 rows × 65 columns

As I observed there are lot or irrelevant features which would not contribute towards enhancing the prediction results.

Hence to get rid of them we use the following code:-

```
data.drop("What is the type of home that you
own?",axis=1,inplace=True)
```

> *Using the .drop() function would help us drop columns which we find unwanted.*

> *axis=1 will ensure that the column is dropped.*

# 2. MANIPULATING THE DATA

The data will be edited to maintain uniformity and consistency. Also, we will fill up the NaN or None value using mean, median or mode.

Missing values or NaN values are present in the dataset when the customer or a layman is used to fill the form for collection of data. This leads to inconsistency in the data which is collected.

- Mean is used when the data is numeric, symmetrical and uniform and does not have any outliers.

- Median is used when data is numeric, left skewed or right skewed and consists of outliers.
- Mode is used when working with categorical data.

We use the following python command to fill NaN or None values, which are also called as missing values:-

```
data["Do you have a home loan? "].fillna(data["Do you have a home
loan? "].mode()[0],inplace=True)
```

- Note that the above code to fill NaN values in the column "Do you have a home loan?" uses mode, as the data present in the column is categorical.

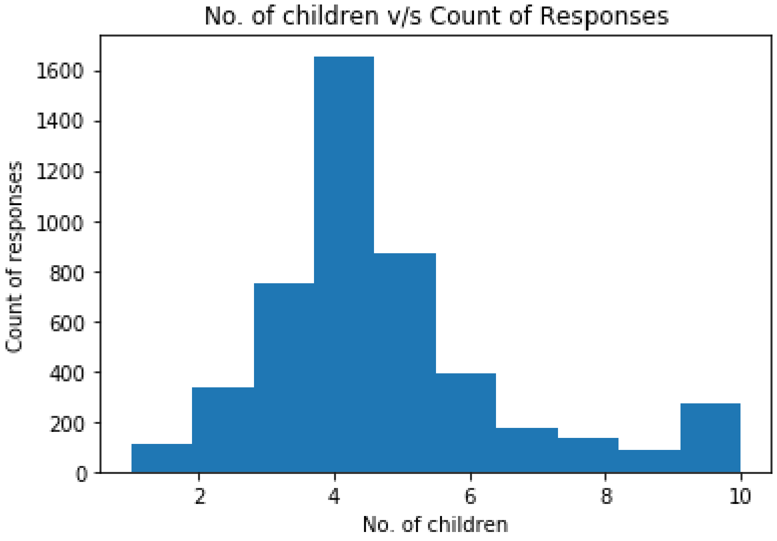
```
data["How many people live in your household(including
children)?"].fillna(data["How many people live in your
household(including children)?"].median(),inplace=True)
```

- Note that the above code to fill NaN values in the column "How many people live in your household(including children)?" uses median, as the data present in the column is numeric, has certain outliers and is slightly left skewed.

> *The skewness and outliers can be detected by the use of a histogram.*

The following code can be used to plot a histogram of the column "How many people live in your household(including children)?"

```
plt.hist(data["How many people live in your household(including
children)?"])

plt.xlabel("No. of children")

plt.ylabel('Count of responses')

plt.title("No. of children v/s Count pf Responses")

plt.show()
```



We will categorize the data of the column "Gender" into three categories:-

* Male

* Female

* Others

```
data['Marital Status']=data['Marital Status'].apply(cleaning_names)
```

Applying an appropriate cleaning_names method which would do the work for us, by categorizing the data into three categories.

# 3. BALANCING THE IMBALANCED DATA

Since the data points are highly imbalanced, for example, the number of customers who purchased the solar panels and the number of customers who did not purchase solar panel was of the ratio 8:139 approximately.

This implies, that for 8 customers who bought the solar panels, there exists 139 customers who did not buy the solar panel. So it becomes extremely crucial for us to balance the data points so that the prediction results won't be biased.

The biased nature of the prediction occurs due to the reason that one value overpowers the other value.

Hence to balance out the imbalanced data we use the following library:-

```
from imblearn.combine import SMOTETomek
```

We plan on over sampling the values. This would result in increased number of data points.

SMOTETomek is a method of imblearn library. It is a hybrid method which uses under sampling method (Tomek) with an over sampling method (SMOTET).

Since the SMOTETomek library uses numerical data to balance out the data points, we have to convert categorical data to numerical data.

Therefore, we use the method:-

```
dummy_columns=pd.get_dummies(categorical_data,drop_first=False)
```

This results in the categorical data to turn into numerical data. And we store those values in dummy_columns.

# 4. SELECTING THE BEST FEATURES

The feature selection step is a vital step in building the prediction model. The features chosen would directly impact the accuracy of the model.

> *Reducing the number of features to be used is beneficial, as large number of features might lead to the 'CURSE OF DIMENSIONALITY'.*

Curse of Dimensionality results in lowering the accuracy of the model.

The Feature Selection technique that we use is: UNIVARITATE FEATURE SELECTION

## Univariate Feature Selection

Statistical tests are used to select the features that have the strongest relationship with the output variable.

Importing the library SelectKBest provides us with a suite of statistical tests to select a specific number of features.

```
from sklearn.feature_selection import SelectKBest
```

From SelectKBest library we wish to use the Chi Square test.

### *CHI SQUARE TEST*

- In feature selection, we aim to select the features which are highly dependent on the response.
- A chi-square test is used in statistics to test the independence of two events. Given the data of two variables, we can get observed count O and expected count E. Chi-Square measures how expected count E and observed count O deviates each other.
- When two features are independent, the observed count is close to the expected count, thus we will have smaller Chi-Square value. So high Chi-Square value indicates that the hypothesis of independence is incorrect. In simple words, higher the Chi-Square value the feature is more dependent on the response and it can be selected for model training.

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

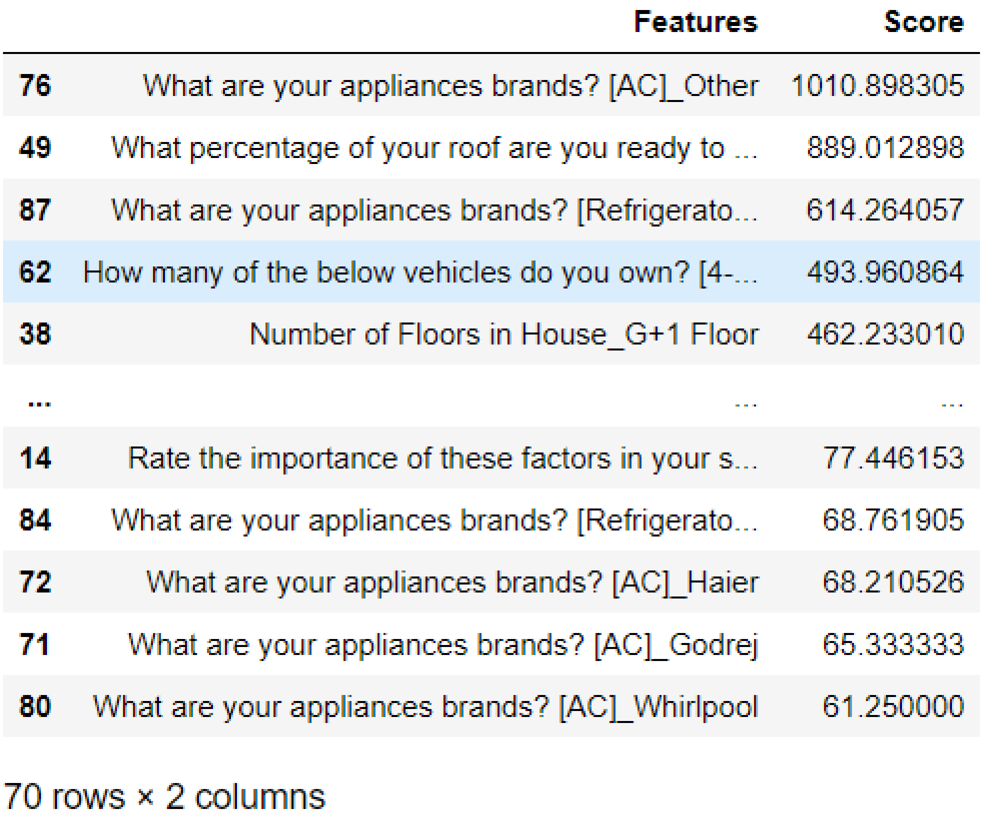
$\chi^2$ = the test statistic $\sum$ = the sum of

O = Observed frequencies E = Expected frequencies

The following library of chi square was used to find the scores for the best feature:-

```
from sklearn.feature_selection import chi2
```

Hence after selecting the top 70 features that would effect the output directly we have the following features:-



| | Features | Score |
|---|---|---|
| 76 | What are your appliances brands? [AC]_Other | 1010.898305 |
| 49 | What percentage of your roof are you ready to ... | 889.012898 |
| 87 | What are your appliances brands? [Refrigerato... | 614.264057 |
| 62 | How many of the below vehicles do you own? [4-... | 493.960864 |
| 38 | Number of Floors in House_G+1 Floor | 462.233010 |
| ... | ... | ... |
| 14 | Rate the importance of these factors in your s... | 77.446153 |
| 84 | What are your appliances brands? [Refrigerato... | 68.761905 |
| 72 | What are your appliances brands? [AC]_Haier | 68.210526 |
| 71 | What are your appliances brands? [AC]_Godrej | 65.333333 |
| 80 | What are your appliances brands? [AC]_Whirlpool | 61.250000 |

70 rows × 2 columns

As we observe the top 3 features  out of the other top features effecting the output are:-

1. What are your appliances brands? [AC]
2. What percentage of your roof are you ready to give for solar?
3. What are your appliances brands? [Refrigerator]

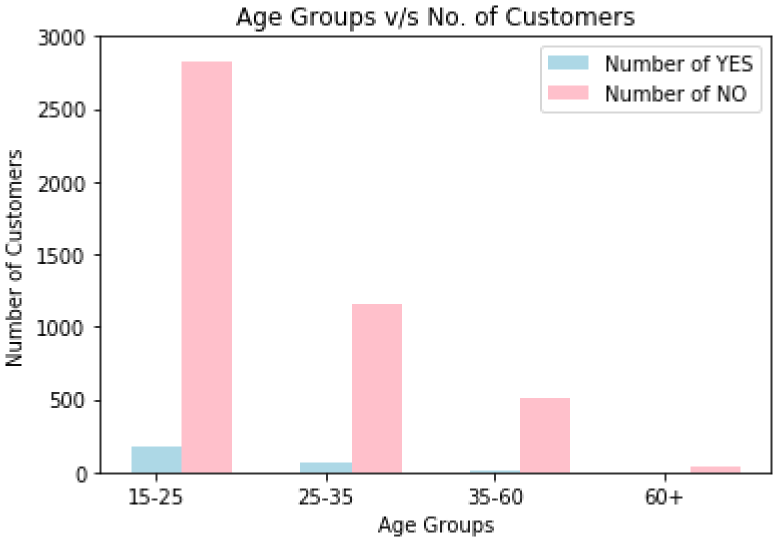
# 5. DATA VISUALIZATION

We must visualize the data by plotting graphs as to see the relationship between the top features that effect the target value or the output, which in our case is to predict whether or not a customer will purchase the solar panel or not.

Observing the data would lead us to create newer insights and would bring out an elaborate view of which section of customers are more drawn towards investing in Solar Panels.

Through this observation, the company can expand their customers in that specific section.

### 5.1. Age Group



ANALYSING THE LEADS WHICH PURCHASED THE SOLAR PANELS [Blue Bins]

---

> *The maximum number of leads are from the age group of people between 15-25.*

> *The second maximum number of leads are from the age group of people between 25-35*

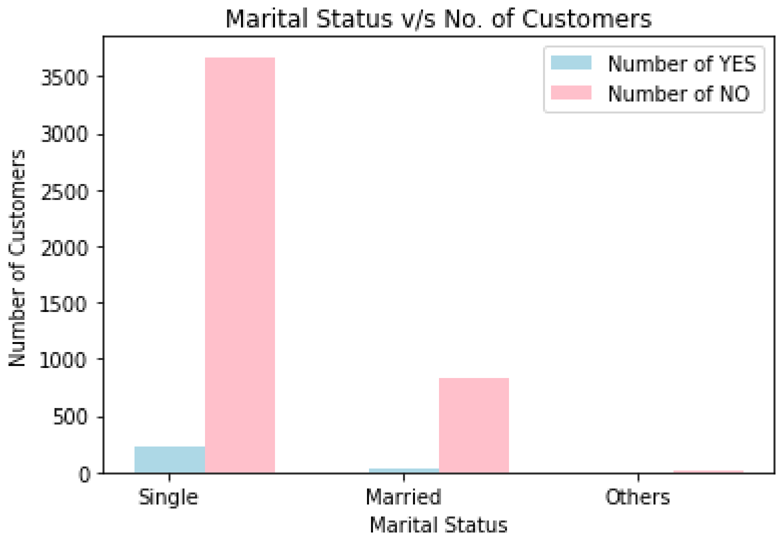
ANALYSING THE LEADS WHICH DID NOT PURCHASED THE SOLAR PANELS [Pink Bins]

---

> *The maximum number of customers who did not purchase are from the age group of people between 15-25.*

> *The second maximum number of customers who did not purchase are from the age group of people between 25-35.*

## 5.2. Marital Status



ANALYSING THE LEADS WHICH PURCHASED THE SOLAR PANELS [Blue Bins]

---

> *The maximum number of leads are having marital status as Single.*

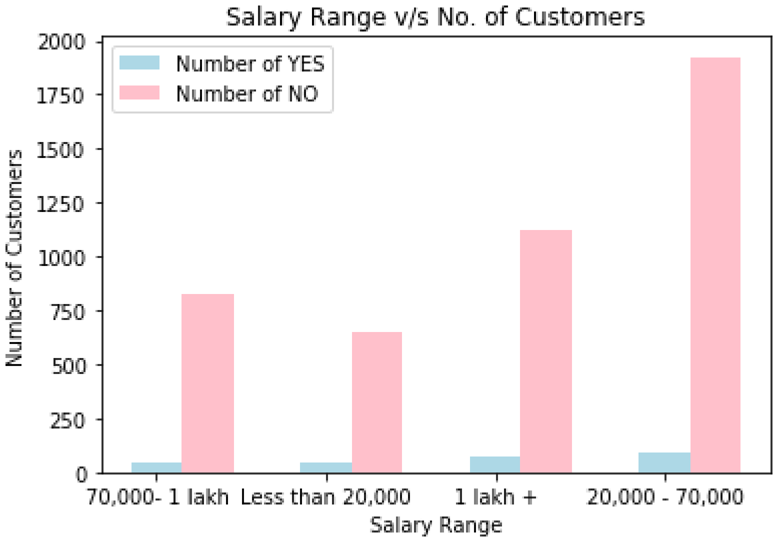
ANALYSING THE LEADS WHICH DID NOT PURCHASED THE SOLAR PANELS [Pink Bins]

---

> *The maximum number of customers who did not purchase are having the marital status as Single.*

> *The second maximum number of customers who did not purchase are having the marital status as Married.*

### 5.3. Salary Range



ANALYSING THE LEADS WHICH PURCHASED THE SOLAR PANELS [Blue Bins]

> *The maximum number of leads are from the salary range of Rs.20,000 - Rs.70,000.*

> *The second maximum number of leads are from the salary range of Rs.1 Lakh+.*

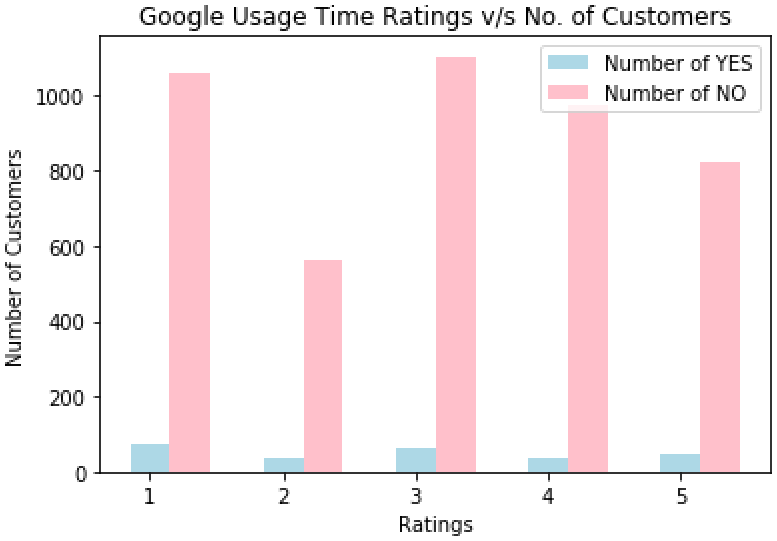
ANALYSING THE LEADS WHICH DID NOT PURCHASED THE SOLAR PANELS [Pink Bins]

> *The maximum number of customers who did not purchase are from the salary range of Rs.20,000 - Rs.70,000.*

> *The second maximum number of customers who did not purchase are from the salary range of Rs.1 Lakh+.*

> *The third maximum number of customers who did not purchase are from the salary range of Rs.70,000 - Rs.1 Lakh+.*

## 5.4. Google Usage Time Ratings



ANALYSING THE LEADS WHICH PURCHASED THE SOLAR PANELS [Blue Bins]

---

> *The maximum number of leads are those who have the minimum Google Usage Time Rating,ie,1.*

> *The second maximum number of leads are those who have Google Usage Time Rating,of 3.*

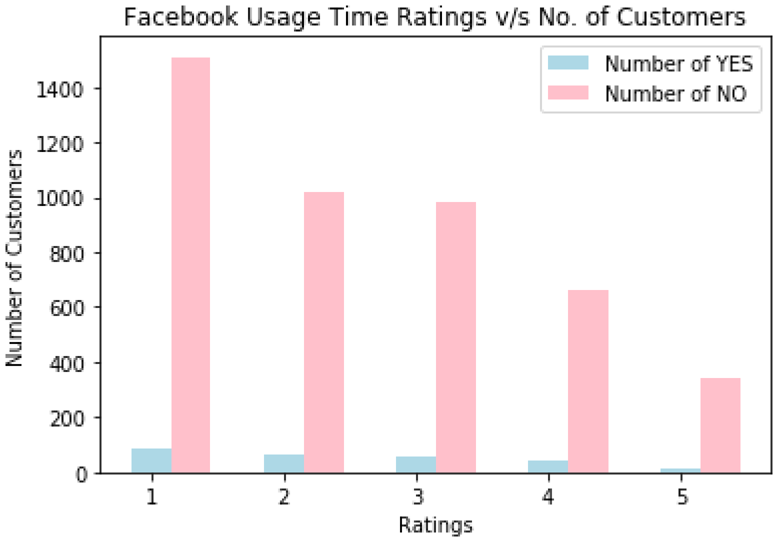
ANALYSING THE LEADS WHICH DID NOT PURCHASED THE SOLAR PANELS [Pink Bins]

---

> *The maximum number of customers who did not purchase are those who have Google Usage Time Rating,of 3.*

> *The second maximum number of customers who did not purchase are those who have the minimum Google Usage Time Rating, i.e, 1.*

### 5.5. Facebook Usage Time Ratings



ANALYSING THE LEADS WHICH PURCHASED THE SOLAR PANELS [Blue Bins]

---

> *The maximum number of leads are those who have the minimum Facebook Usage Time Rating of 1.*

> *The second maximum number of leads are those who have Facebook Usage Time Rating of 2 and 3.*

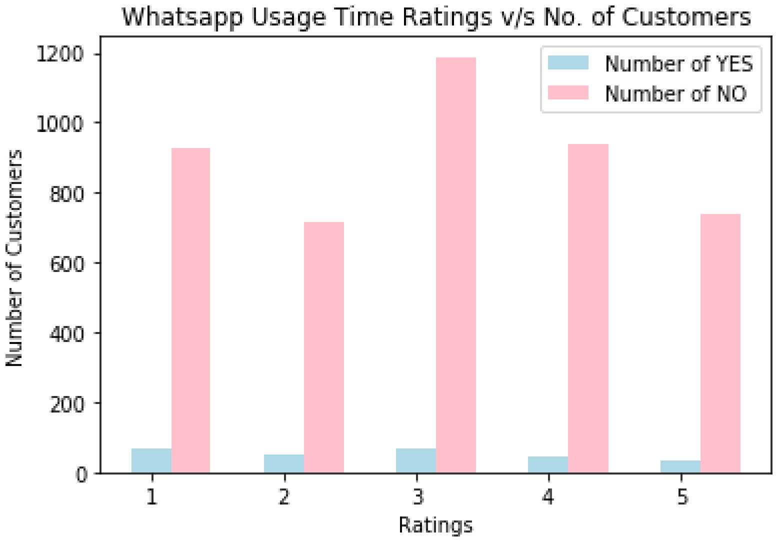
ANALYSING THE LEADS WHICH DID NOT PURCHASED THE SOLAR PANELS [Pink Bins]

---

> *The maximum number of customers who did not purchase are those who have Facebook Usage Time Rating of 1.*

> *The second maximum number of customers who did not purchase are those who have the Facebook Usage Time Rating of 2.*

## 5.6. Whatsapp Usage Time Ratings



ANALYSING THE LEADS WHICH PURCHASED THE SOLAR PANELS [Blue Bins]

---

> *The maximum number of leads are those who have the Whatsapp Usage Time Rating of 1 and 3.*

> *The second maximum number of leads are those who have Whatsapp Usage Time Rating of 2.*

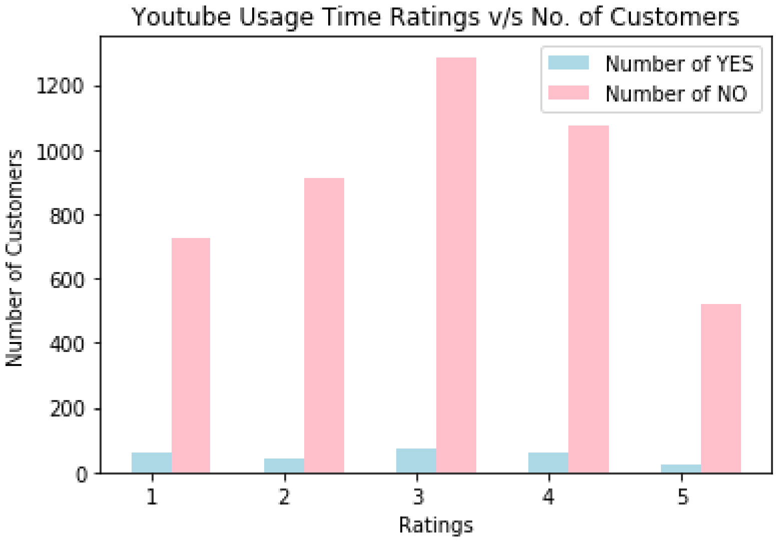
ANALYSING THE LEADS WHICH DID NOT PURCHASED THE SOLAR PANELS [Pink Bins]

---

> *The maximum number of customers who did not purchase are those who have Whatsapp Usage Time Rating of 3.*

> *The second maximum number of customers who did not purchase are those who have the Facebook Usage Time Rating of 1 and 4.*

### 5.7. YouTube Usage Time Ratings



ANALYSING THE LEADS WHICH PURCHASED THE SOLAR PANELS [Blue Bins]

> *The maximum number of leads are those who have the minimum YouTube Usage Time Rating of 3.*

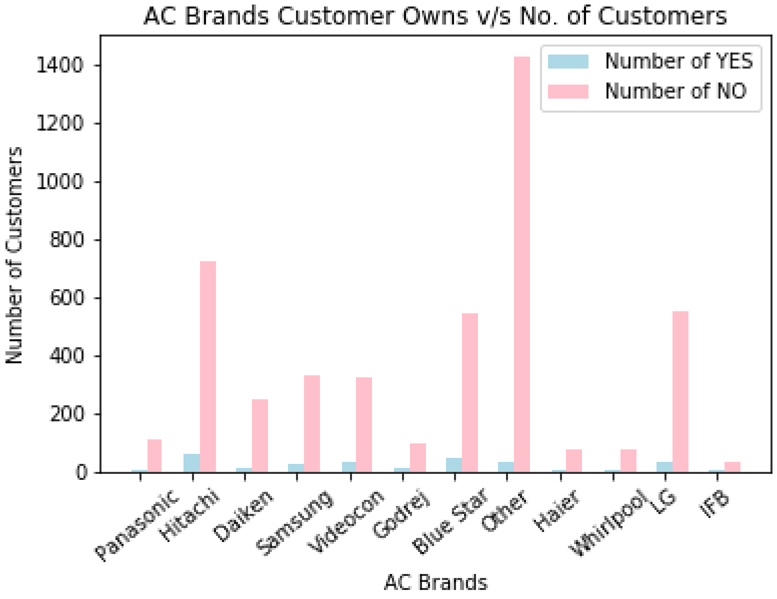
> *The second maximum number of leads are those who have YouTube Usage Time Rating of 1 and 4.*

ANALYSING THE LEADS WHICH DID NOT PURCHASED THE SOLAR PANELS [Pink Bins]

> *The maximum number of customers who did not purchase are those who have YouTube Usage Time Rating of 3.*

> *The second maximum number of customers who did not purchase are those who have the YouTube Usage Time Rating of 4.*

## 5.8. Air Conditioner Brand Owned by the Customer



ANALYSING THE LEADS WHICH PURCHASED THE SOLAR PANELS [Blue Bins]

> *The maximum number of leads are from the customers who owns air conditioner from the brand Hitachi.*

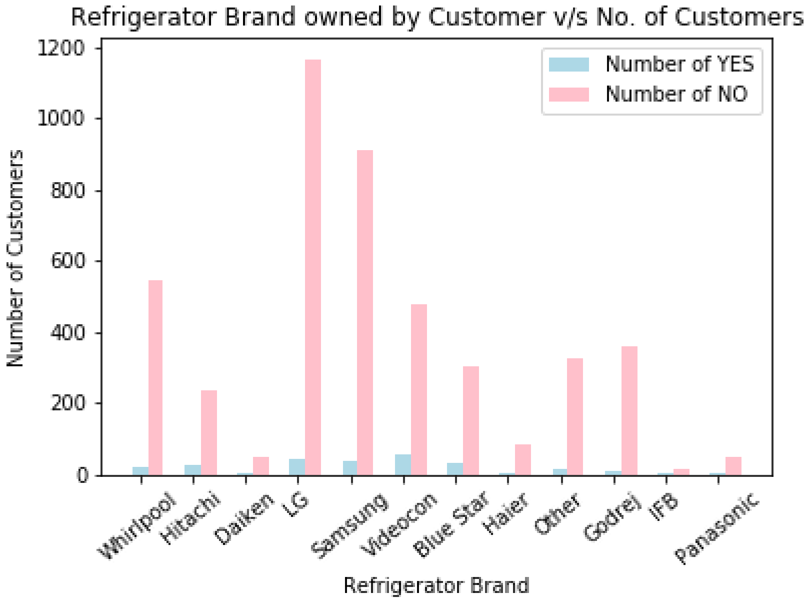
> *The second maximum number of leads are from the customers who owns air conditioner from the brand Blue Star.*

ANALYSING THE LEADS WHICH DID NOT PURCHASED THE SOLAR PANELS [Pink Bins]

> *The maximum number of customers who did not purchase are from the customers who owns air conditioner from other than the brands mentioned in the survey.*

> *The second maximum number of customers who did not purchase are from the customers who owns air conditioner from the brand Hitachi.*

## 5.9. Refrigerator Brand Owned by the Customer



ANALYSING THE LEADS WHICH PURCHASED THE SOLAR PANELS [Blue Bins]

---

> *The maximum number of leads are from the customers who owns refrigerator from the brand Videocon.*

> *The second maximum number of leads are from the customers who owns refrigerator from the brand LG.*

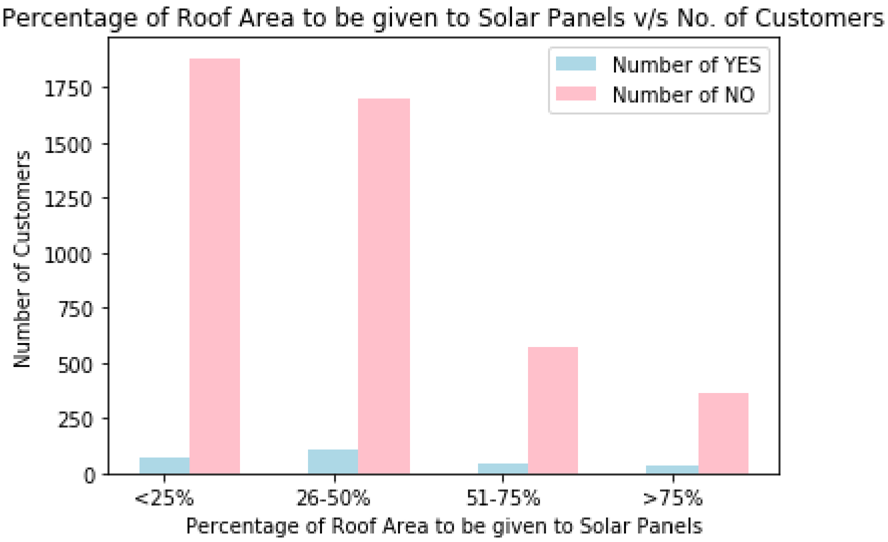
ANALYSING THE LEADS WHICH DID NOT PURCHASED THE SOLAR PANELS [Pink Bins]

---

> *The maximum number of customers who did not purchase are from the customers who owns refrigerator from the brands LG.*

> *The second maximum number of customers who did not purchase are from the customers who owns refrigerator from the brand Samsung.*

## 5.10. The Percentage of Roof Area that Customer is Willing to Provide for Solar Panels



ANALYSING THE LEADS WHICH PURCHASED THE SOLAR PANELS [Blue Bins]

---

> *The maximum number of leads are from the category who would provide 26% – 50% of their roof area for the installation of Solar Panels.*

> *The second maximum number of leads are from the category who would provide less than 25% of their roof area for the installation of Solar Panels.*

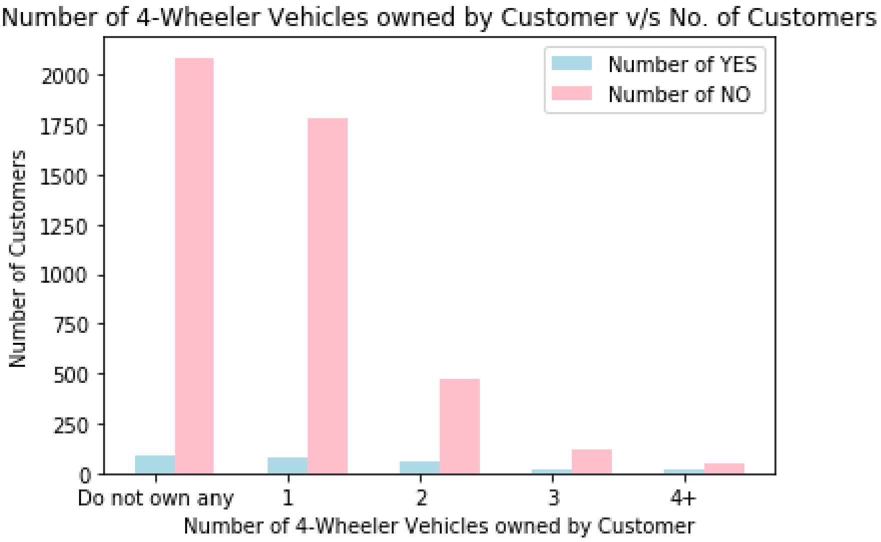
ANALYSING THE LEADS WHICH DID NOT PURCHASED THE SOLAR PANELS [Pink Bins]

---

> *The maximum number of customers who did not purchase are from the category who would provide less than 25% of their roof area for the installation of Solar Panels.*

> *The second maximum number of customers who did not purchase are from the category who would provide 26% – 50% of their roof area for the installation of Solar Panels.*

### 5.11. Number of 4-Wheeler Vehicles Owned by the Customers



ANALYSING THE LEADS WHICH PURCHASED THE SOLAR PANELS [Blue Bins]

---

> *The maximum number of leads are from the category where the customers do not own any 4-wheeler vehicle.*

> *The second maximum number of leads are from the category where the customers own 1, 4-wheeler vehicle.*

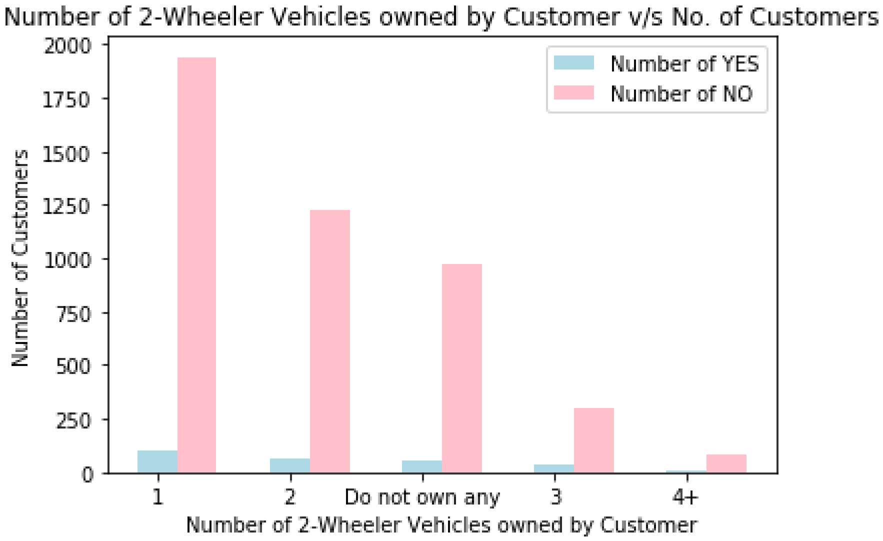
ANALYSING THE LEADS WHICH DID NOT PURCHASED THE SOLAR PANELS [Pink Bins]

---

> *The maximum number of customers who did not purchase are from the category where the customers do not own any 4-wheeler vehicle.*

> *The second maximum number of customers who did not purchase are from the category where the customers own 1, 4-wheeler vehicle.*

### 5.12. Number of 2-Wheeler Vehicles Owned by the Customer



ANALYSING THE LEADS WHICH PURCHASED THE SOLAR PANELS [Blue Bins]

---

> *The maximum number of leads are from the category where the customers own 1, 2-wheeler vehicle.*

> *The second maximum number of leads are from the category where the customers own 2, 2-wheeler vehicle.*

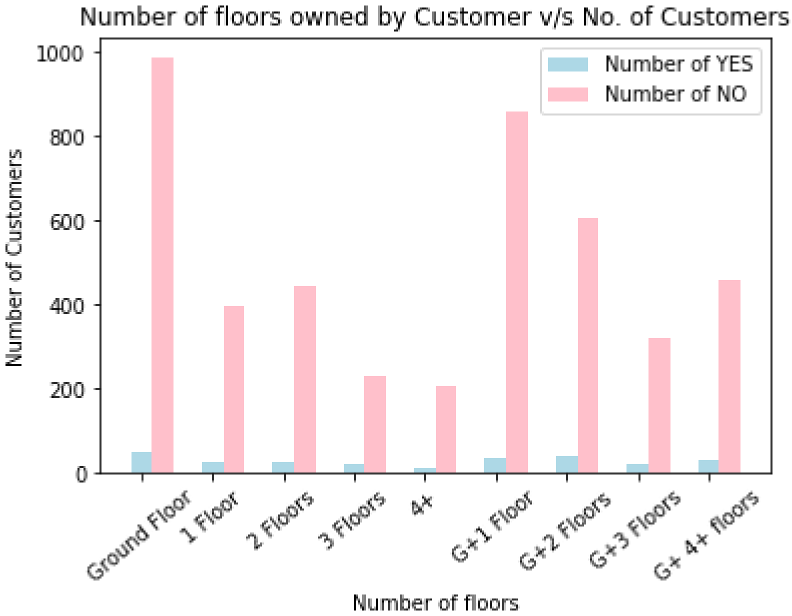
ANALYSING THE LEADS WHICH DID NOT PURCHASED THE SOLAR PANELS [Pink Bins]

---

> *The maximum number of customers who did not purchase re from the category where the customers own 1, 2-wheeler vehicle.*

> *The second maximum number of customers who did not purchase are from the category where the customers own 2, 2-wheeler vehicle.*

### 5.13. Number of Floors the Owner Owns



ANALYSING THE LEADS WHICH PURCHASED THE SOLAR PANELS [Blue Bins]

---

> *The maximum number of leads are from the category where the customer owns only the ground floor.*

> *The second maximum number of leads are from the category where the customer owns only the ground with first floor and ground floor with first & second floors.*

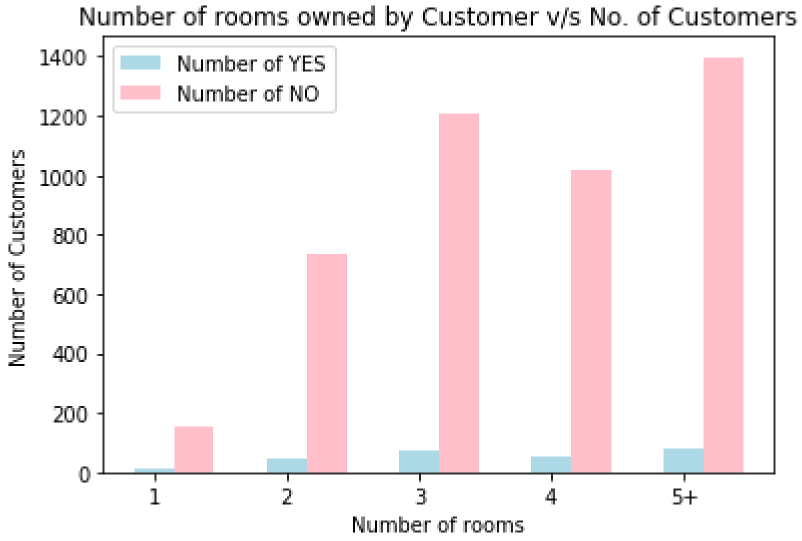
ANALYSING THE LEADS WHICH DID NOT PURCHASED THE SOLAR PANELS [Pink Bins]

---

> *The maximum number of customers who did not purchase are from the category where the customer owns only the ground floor.*

> *The second maximum number of customers who did not purchase are from the category where the customer owns only the ground with first floor.*

### 5.14. Number of Rooms the Owner Owns



ANALYSING THE LEADS WHICH PURCHASED THE SOLAR PANELS [Blue Bins]

---

> *The maximum number of leads are from the category where the customer owns 5+ rooms.*

> *The second maximum number of leads are from the category where the customer owns 3 rooms.*

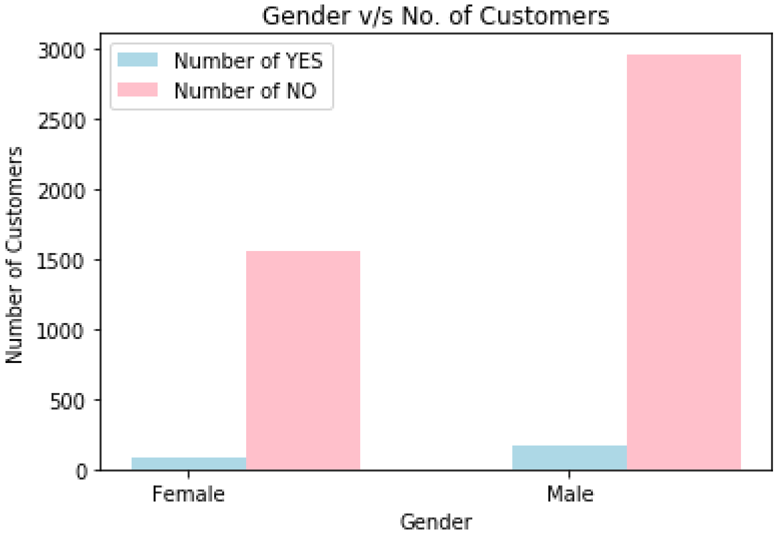
ANALYSING THE LEADS WHICH DID NOT PURCHASED THE SOLAR PANELS [Pink Bins]

---

> *The maximum number of customers who did not purchase are from the category where the customer owns 5+ rooms.*

> *The second maximum number of customers who did not purchase are from the category where the customer owns 3 rooms.*

### 5.15. Gender



ANALYSING THE LEADS WHICH PURCHASED THE SOLAR PANELS [Blue Bins]

---

> *The maximum number of leads are from customers who are Male.*

ANALYSING THE LEADS WHICH DID NOT PURCHASED THE SOLAR PANELS [Pink Bins]

---

> *The maximum number of customers who did not purchase are Male.*

## ANALYSING THE STATES WHERE SOLAR PANELS WERE NOT SOLD

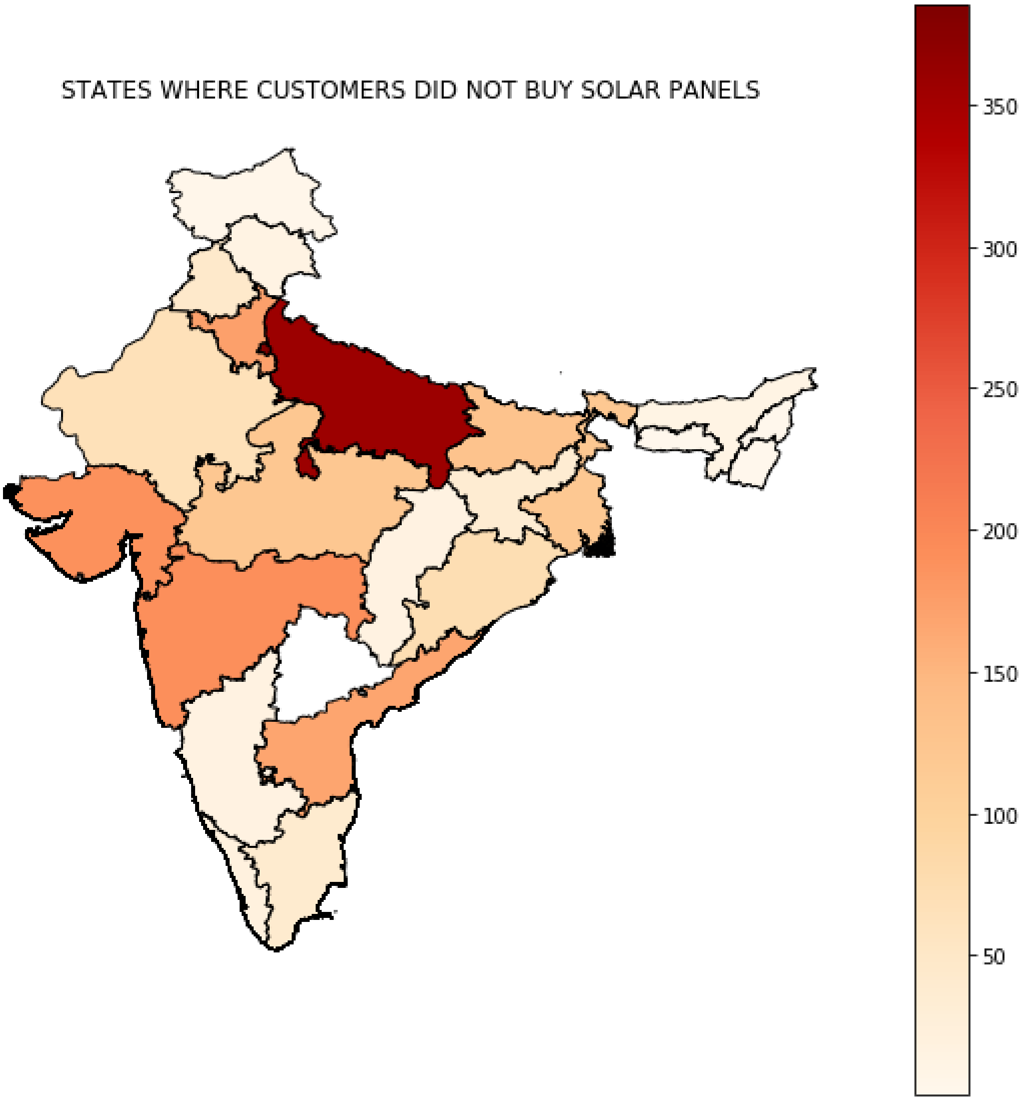
By analyzing the states where solar panels are not sold we could understand what aspects are lacking in sales.

To expand the sales of solar panel we must evaluate and analyze the states where the solar panels were not sold so that, the sales department could evaluate the reason and enhance the sales strategy.

We use the following library to visualize data:-

```
import geopandas as gpd
```

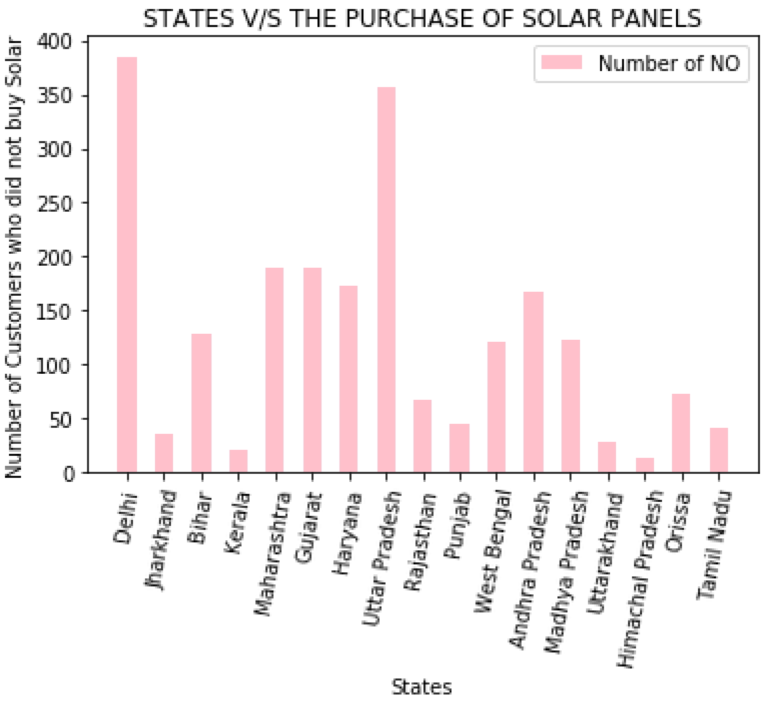
Now, using the map of India we would visualize the data where there were no sales.



> *The deeper the colour of red, least are the sales in the specific state.*

Through the map it is clear that the least sales record is established in the state of Delhi and Uttar Pradesh, followed by Gujarat and Maharashtra.

For further analysis, we must refer to the following bar chart to see the performance of sales in other states as well.

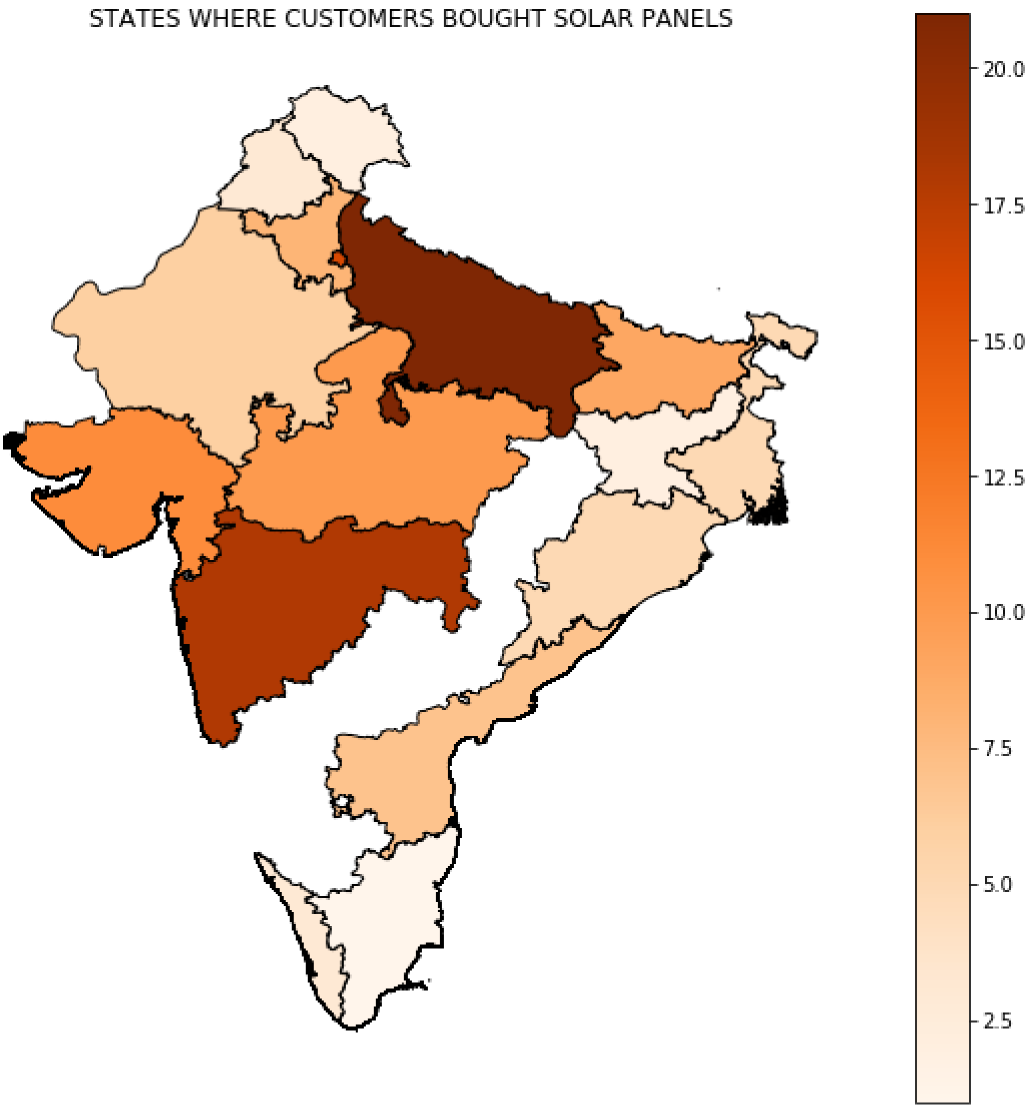


## ANALYSING THE STATES WHERE SOLAR PANELS WERE SOLD

To expand the sales of solar panel we must also evaluate and analyze the states where the solar panels were sold the maximum, so that, expansion could be done in those sectors.

This would help generate great revenue for the company as the data suggests the maximum sales in the following states.

Now, using the map of India we would visualize the data where the sales were maximum.

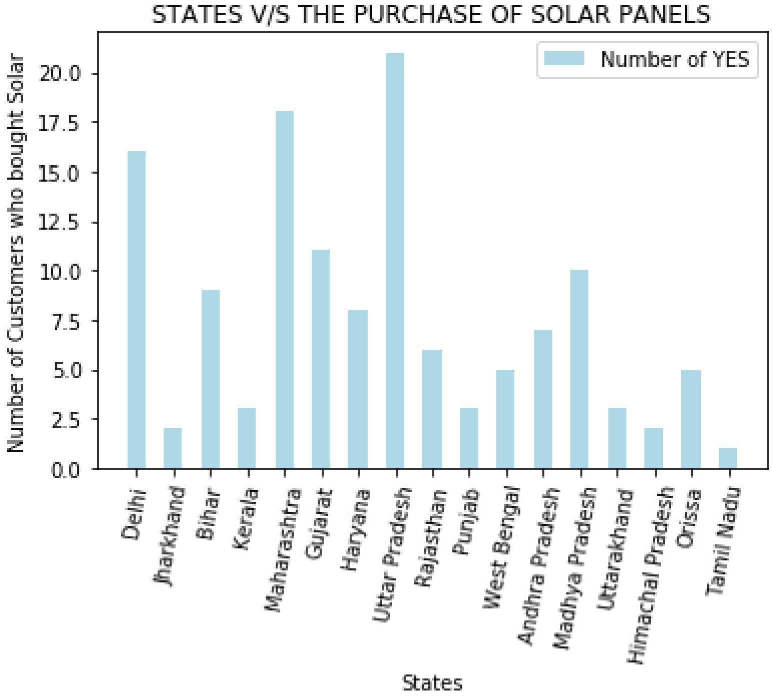


> *The deeper the colour of red, more the sales in the specific state.*

> *The lighter the colour of red, lesser the sales in the specific state.*

Here, via the map it is clear that the top sales record is established in the state of Uttar Pradesh, followed by Maharashtra and Delhi.

For further analysis, we must refer to the following bar chart to see the performance of sales in other states as well.



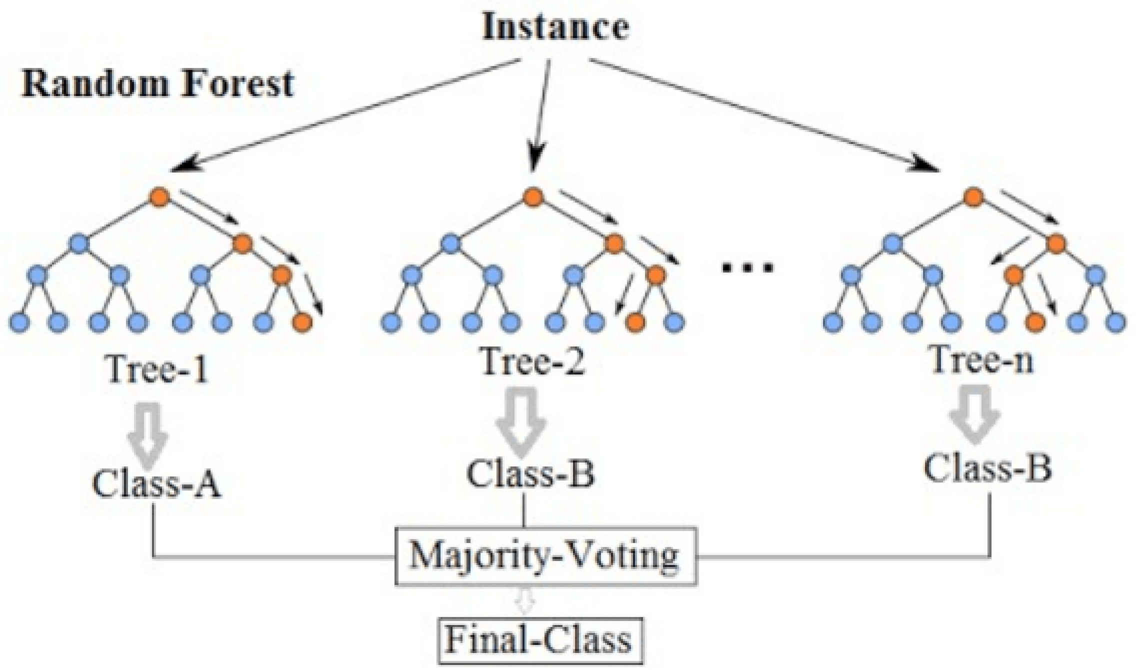
# 6. DATA PREDICTION AND MODEL CREATION

For predicting whether or not the customer would purchase the solar panel, we would perform binary classification.

Hence, we use Random Forest Classification to obtain best results.

### Random Forest Classifier

It is an ensemble tree-based learning algorithm. The Random Forest Classifier is a set of decision trees from randomly selected subset of training set. It aggregates the votes from different decision trees to decide the final class of the test object.



We first import the library to use Random Forest Classifier.

```
from sklearn.ensemble import RandomForestClassifier
```

Also, we would split the data into training data and testing data. So that we could train our classifier and then test the accuracy of the classifier.

> *This is to avoid overfitting and underfitting.*

Importing the library to split the data and splitting the data:-

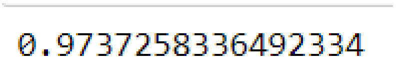
```
from sklearn import model_selection

x_train,x_test,y_train,y_test=model_selection.train_test_split(final_x
_data,final_target,random_state=1)
```

We use the cross validation score to determine the training score.

```
from sklearn.model_selection import cross_val_score
```

**The Training Score obtained:-**

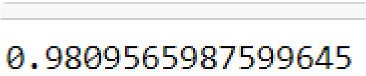
```
cross_val.mean()
```



```
0.9737258336492234
```

**The Testing Score obtained:-**

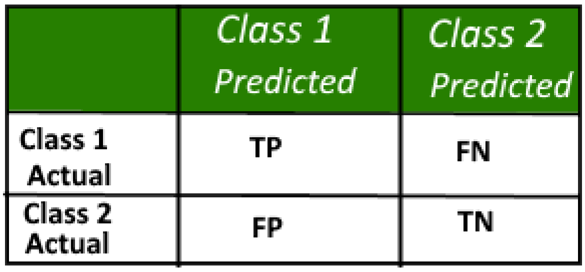
```
clf.score(x_test,y_test)
```



```
0.9809565987599645
```

### *CONFUSION MATRIX* and *CLASSIFICATION REPORT*

> *A **Confusion Matrix** is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm.*



| | Class 1 Predicted | Class 2 Predicted |
|---|---|---|
| Class 1 Actual | TP | FN |
| Class 2 Actual | FP | TN |

The definition of the terms are:-

Class 1 : Positive

Class 2 : Negative

Positive (P) : Observation is positive (for example: is an apple).

Negative (N) : Observation is not positive (for example: is not an apple).

True Positive (TP) : Observation is positive, and is predicted to be positive.

False Negative (FN) : Observation is positive, but is predicted negative.

True Negative (TN) : Observation is negative, and is predicted to be negative.

False Positive (FP) : Observation is negative, but is predicted positive.

> *A **Classification report** is used to measure the quality of predictions from a classification algorithm. How many predictions are True and how many are False. More specifically, True Positives, False Positives, True negatives and False Negatives are used to predict the metrics of a classification report as shown below.*

The definition of terms are:-

1. Precision – Accuracy of positive predictions.

   Precision = TP/(TP + FP)

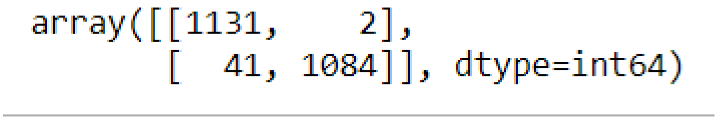
2. Recall: Fraction of positives that were correctly identified.

   Recall = TP/(TP+FN)

3. F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0

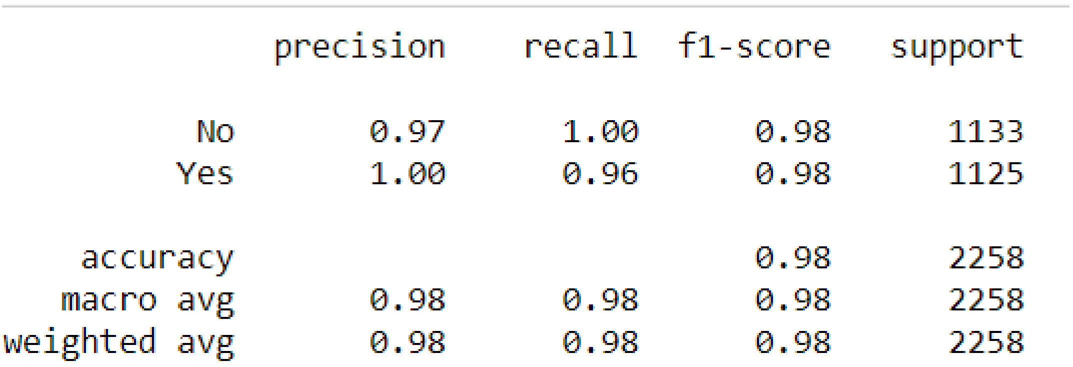
   F1 Score = 2*(Recall * Precision) / (Recall + Precision)

### In our case the CONFUSION MATRIX yields the following results:-



```
array([[1131,    2],
       [  41, 1084]], dtype=int64)
```

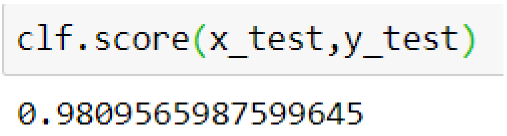
### In our case the CLASSIFICATION REPORT generated the following results:-



```
              precision    recall  f1-score   support

          No       0.97      1.00      0.98      1133
         Yes       1.00      0.96      0.98      1125

    accuracy                           0.98      2258
   macro avg       0.98      0.98      0.98      2258
weighted avg       0.98      0.98      0.98      2258
```

- 'Yes' represents the customers who bought solar panels
- 'No' represents the customers who did not buy solar panels

> ***THE ACCURACY GENERATED BY THE MODEL IS:-***



```
clf.score(x_test,y_test)
```

```
0.9809565987599645
```

- The accuracy obtained is 98.09%

# 7. CONCLUSION

1. The states for expansion as seen from above graph under the section **"ANALYSING THE STATES WHERE SOLAR PANELS WERE SOLD"** we conclude that the states where the sales could be expanded are:- ***UTTAR PRADESH***, ***DELHI***, ***MAHARASHTRA***, ***MADHYA PRADESH***, ***ANDHRA PRADESH***.

2. The classification model prepared by us give the test accuracy of ***98.09%***.