# CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# EXECUTIVE SUMMARY

This project is about the planning activities involved in the summer internship project at Peacock Solar, Gurgaon, Haryana. The report provides with the details of the processes involved in the creation of the predictive model. The report focusses on how the Predicting the customer purchase behavior of home solar panels: a case study of Peacock Solar Ltd is created.

India is increasingly shifting towards solar power in the past few years, with large projects being completed around the country some going as high as 12.2GW. Latest advancements in the solar field has enabled to make the solar cells more power generating as well as long lasting with life span between 20 to 25 years. Peacock solar is a yet another company started with the dream to convert as many households as possible to solar power dependent. Unlike other companies Peacock solar utilizes IoTs, and Data science to better identify the needs and areas where the solar panels can be better suited. Their focus is mainly towards rooftop solar panel installation, which can not only provide power needs of every household but also creating an earning opportunity from empty roof with their grid connected solar setup. They also provide maintenance services as well as helps the customer to get subsidy from state government based of the polices existing in that particular state. Since the initial cost of solar system installation is very high and many customers don’t install them due to lack of this initial funds, Peacock solar provides EMI schemes for those having difficulty in lump sum payment with flexible tenures.

# CHAPTER – 1
# INTRODUCTION

## 1.1 Introduction About the Study:

In this study we are trying to create a prediction model with different classification algorithm to identify if a customer is likely to purchase a home solar panel based on the various independent variables. The customer purchase interest column in the dataset is taken as the dependent variable, and rest of the attributes are taken as the independent variable or predictor variables.

**Steps involved in the study:**

- ✓ Cleaning the data and removal of null values if any
- ✓ Visualization of data to understand it better
- ✓ Balancing of imbalanced training dataset
- ✓ Training different classification models
- ✓ Cross-validation of the model to ensure better fit
- ✓ Creating confusion matrix to determine the predication performance.

### 1.1.1 Solar Power:

Solar energy is one of the natural occurring abundant source of energy which can be used is various. Two of the most common uses of the Solar energy is using to convert the heat energy from solar to boiling of water, commonly known as solar water heaters. Another major way the solar energy is utilized is to convert this thermal energy of sun light to electrical power which made possible with the help of Photovoltaic cell (PV cells), This cell absorbs sun light and convert the thermal energy from sun light in to electrical power.

Combination of these cells are used to generate different volts and watts of power. The power generated from the PV cells are DC power, which is then converted to AC power of required volts using transformers. Solar power is one of the best renewal energies that is abundant in our planet. But the use of Solar cells depends on different countries, their power generation cost, material cost etc.

### 1.1.2 Power Generation Capacity:

A 1kw Solar panel can generate anywhere between 1400 to 1600 units of electricity per year and has a life span of 20 to 25 years. That about 4 units of electricity generated every day, in places where abundance of Sun light is present. Solar power generation efficiency is based on the amount of sun light present, So the solar panels needs to be placed at an area when there is not obstruction of sunlight is present.

### 1.1.3 **Raw Materials Used to Make Solar Panels:**

- Sand, silicon, ingots, wafers, and, solar cells are used for manufacturing the solar panels.
- Silicon is found in the sand, mostly in natural beach sand, which is richly available.
- Conversion of sand into silicon is the most important and primary step.
- Ingots are cylindrically shaped, melted compound which we get from silicon rocks.
- After ingots are carved into thin disks, we get wafers.
- Solar cells are formed by covering wafers with metal conductors capable of seizing solar rays and changing them into electricity, and then solar cells are combined together to form a matrix like assembly called solar panels

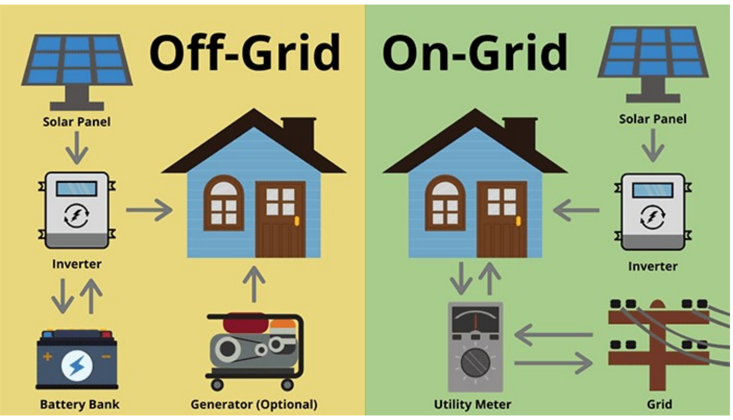
### 1.1.4 **Types of Solar Rooftops:**

- Monocrystalline –These cells are cut out of an ingot grown from a single large crystal of silicon. These kinds of panels are more expensive and space wise they very efficient.
- Polycrystalline – these cells are cut from an ingot made of many small crystals of silicon. These panels are less expensive and have slightly lower heat tolerance.
- Amorphous – are a thin small solar cell. They are very thin, small panels made of several layers of photovoltaic material.

### 1.1.5 **Types of Solar Installations:**

The solar panels are installed in houses and buildings in two different ways:

- ON-Grid Solar
- Off-Grid Solar



*Source: paradisesolarenergy.com*

***Figure - 1***

**On-Grid Solar** installation involves the setup of solar panels on roof top or suitable places directly connected with power grid that is coming to the house of building. The main advantage of this type of solar system is that no power storage devices, i.e. batteries are needed to storage the extra power generated by the solar panel, and when solar energy is not available the customers can use the power coming from the power grid automatically.

Another advantage of this setup is that the excess power generated by the solar panel is sent to the central power grid and the government will provide you with money for that excess power.

The disadvantage of using this setup is that the solar panel keep on working on when there is power coming from the central power grid, if for some reason no electricity is coming from power grid the solar panel will stop generating electricity, this is a safety mechanism done to prevent damage of applicable due to deviated power generation by solar panel in this On Grid setup.

**Off-Grid** Solar installation, as the name indicates the installation enables the house or building to be completely off grid from the central power grid, this means that the house or building where off-grid solar panels are installed can run completely of the solar power 24/7.

The main advantage of this system is that that the power generated by the solar panels is readily available for use and excess power generated is stored in batteries which can be utilized when solar power generation is not available. This type of setup would require more KW solar panels as the sole source of power the customer is from solar panel only,

A sufficiently large battery is also required to store the excess power generated. The drawback of this setup is that the battery needs regular maintenance and unlike solar panels, the battery has less life span and most probably needs to be replaced within 4 to 6 years. Hence cost associated with off grid solar panel installation is comparatively on the higher end.

## 1.2 Company Profile

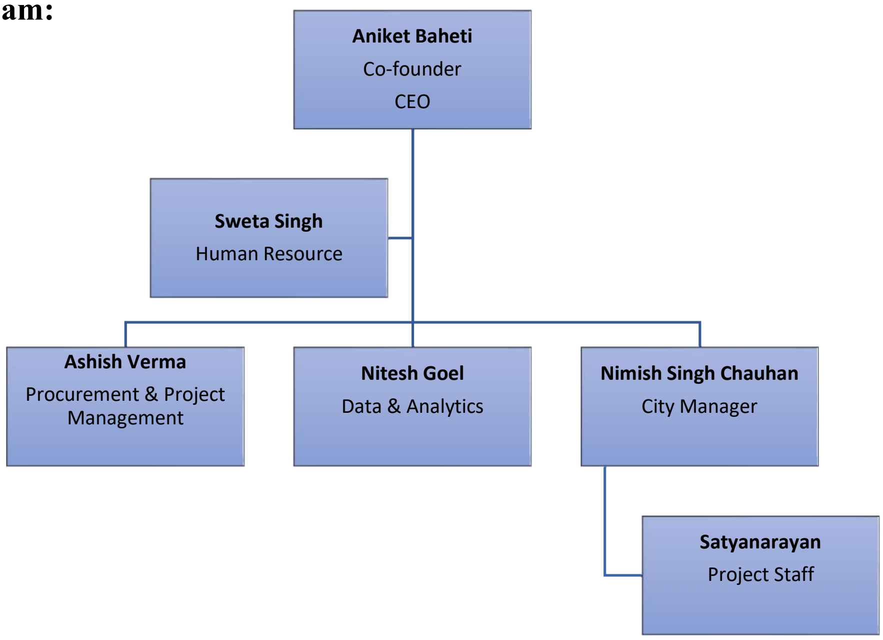
### Peacock Solar Group:

Peacock solar was founded in the year 2017, the CEO of the company is Aniket Baheti a solar professional since 2010 and brings in experience managing diverse operations. He is an alumni of IIT Madras and Indian School of Business. Peacock Solar-A residential rooftop solar company, aspiring to make an impact on the community. Their aim is to provide zero-maintenance energy solutions to convert idle rooftops to sustainable powerhouses with premium and reliable materials.

Peacock Solar was founded with the vision to empower India's 50 million households with access to clean renewable energy. Peacock solar utiilizes data analytics and innovative finance to make solar panel installations more affordable and reliable for house owners across the country. They offer superior quality of technical know-how in their end-to-end suite of solar panel installation services. They are honored to be recognized by Climate Finance Lab as one of the top 9 global ideas for sustainable development in 2018 cycles and they are funded by UNICEF to further their vision of getting solar at every home.

### 1.2.1 Organization Structure:

**Core Team:**



## 1.3 Company Details:

| | |
|---|---|
| Website | https://peacock.solar/ |
| Company type | Privately held |
| Company size | 10 -50 |
| Headquarters | Gurgaon, Haryana |
| Year founded | 2017 |
| Specialties | Rooftop solar panel, installation, maintenance |

# CHAPTER – 2

# REVIEW OF LITERATURE

Eunju Kim, Wooju Kim & Yillbyung Lee (2002), E-business businesses are keen to understand about their consumers by employing data mining technologies. But the distinct situations of such companies make it difficult to know which is the most effective algorithm for the given issues. Recently, a development towards combining multiple classifiers has emerged to increase classification results. In this study, they have suggested an method for the prediction of the E-commerce customer's purchase behavior by utilizing multiple classifiers based on genetic algorithm. The method was tested and evaluated using Web data from a leading E-commerce company. They have also tested the validity of their approach in general classification problems using handwritten numerals. In both cases, their method produces better performance than individual classifiers and other known combining methods.

Chelsea Schelly (2004), This study analyzes the question, what motivates homeowners to adopt residential solar electric technology? They conducted through interviews with 48 people across the state of Wisconsin, this study explores the relative importance environmental reasons, economic considerations, and the demographic characteristics and network relations affecting the adoption and diffusion of innovations. This research indicates (1) environmental values alone are not sufficient, and are not always necessary, to drive adoption; (2) rational economic calculation in the narrow sense of calculated return on investment or payback time is less significant than the specific timing of economic events within a household; and (3) perceiving oneself as an early adopter is particularly important for some consumers .these Wisconsin homeowners shared an unexpected characteristic that they identified as motivating adoption – a passion in technical innovation and enjoyment of the technical aspects of power systems. The discoveries from this empirical case study offer general insight for understanding investment in renewable energy technologies at the residential scale, indicating means of promoting environmental and energy policy and highlighting avenues for subsequent research.

Serhat Peker, Altan Kocyigit and P. Erhan Eren, (2017), Predicting customers' purchase interest is a very challenging task. The research has introduced the individual-level and the segment-based predictive modeling approaches for this purpose. Each approach has its own advantages and drawbacks, and performs in certain cases. The reason of this study is to come up with a hybrid technique which predicts consumers' individual purchase interest and reduces the limitations of these two methods by incorporating the advantages of them.

The proposed hybrid technique is formulated based on individual level and segment-based approaches and utilizes the historical transactional data and predictive algorithms to develop predictions. The effectiveness of the proposed method is experimentally calculated in the field of supermarket shopping by utilizing real-world data and implementing five popular machine learning classification algorithms including logistic regression, random forests, decision trees, neural networks and support vector machines

C.Robert Newberry,Bruce R. Klemz, & Christo Boshoff(2003), Services managers are dependent on forecasting purchase interest when creating resource allocation choices. Purchase intentions are frequently used as a base to forecast purchase behavior. This method is, however, not without its critics. In a study of restaurant patrons, it was discovered that the patrons who showed strong purchase intent and carried out a subsequent purchase demonstrated unique attitude differences when related to those patrons who also showed strong purchase intent but neglected to make a subsequent purchase. The results indicate that the service manager could be mistaken, and therefore could make costly service mix mistakes, if purchase intent is employed solely to model purchase behavior.

P. Anitha & Malini M. Patil (2019), The purpose of this research is to implement business intelligence in spotting potential consumers by provid-ing relevant and timely data to business entities in the Retail Industry. The data proccured is based on systematic research and scientific application in evaluation of the sales history and purchasing interest of the customers. The curated and organized data as a result of this scientific research not only enhances busi-ness sales and profit, but also supplies with intelligent insights in forecasting consumer purchasing interest and associated patterns. In order to deal with and implement the scientific approach using K-Means algorithm, the real time transactional and retail dataset are considered. Spread over a specific period of business trans-actions, the dataset values and guidelines provide an organized means of the customer purchasing patterns and behavior across different regions. This research is based on the Recency, Frequency and Monetary model and implements dataset segmentation principles using K-Means/K-NN Algorithm. A mixture of dataset clusters are validated based on the estimation of Silhouette Coefficient. The results thus obtained with respect to sales transactions are analyzed with various parameters like, Sales Volume, Sales Frequency and Sales Recency

Haifeng Zhang, Yevgeniy Vorobeychik, Joshua Letchford and Kiran Lakkaraju(2014), this study they have presented is in a novel agent-based modeling methodology to predict rooftop solar adoptions in the residential power market. They first incorporated different linear regression models to predict missing variables for non-adopters, this is done so as to find that attributes of non-adopters and adopters could be used to train a logistic regression model. Then, they combined the logistic regression model along with other predictive models into a multi-agent simulation setup and cross-validated their prediction models by correlating the forecast of aggregate adoptions in a certain zip code area. This result showed that the agent-based model can accurately predict future adoptions. Finally, based on the cross validated agent-based model.

# CHAPTER - 3

# METHODOLOGY

This chapter explains about the methodology adopted for the study, which includes research approach, setting for the study. Population, sample and criteria, Data balancing technique used, selection, method of data collection involved.

Research can be interpreted as the formulation of new observation or the use of existing knowledge in a different and innovative way so as to develop new concepts, methodologies and understandings. This could consist of synthesis and analysis of previous research to the degree that it serves to new and creative outcomes.

## Objective of The Study:

The study aims to improve the Operational efficiency of the peacock solar sales team with the help of predictive model on customer interest in solar panel purchase. The project deals with the creation of a Predicting the customer purchase Behavior of home solar panels: a case study of Peacock Solar Ltd.

This study aims to help the sales team to better analyze customer interest in purchasing solar panels based on previous collected data from these customers. This model allows the sales to get a better probable customer list who are likely to be interested in solar panel purchase and hence only contact this customer and improve their working efficiency and deal closing chances.

- **Research Design:**

  Experimental research:

  In general, experimental research is a type of classical scientific experiment, this research is conducted with a scientific approach using two set of variables. The independent variables are manipulated and applied to the dependent variables to measure their effect on the dependent variable. The effect of independent variable on the dependent variable is measured and recorded to aid the researchers in drawing a conclusion.

- **Research Population involved:**

  Around 2000 customers data collected from across the country involving big cities of different states.

- **Data collected and used for Analysis:**

  Primary data sources: The data used here for the analysis is collected from the company.

  Secondary Sources: Secondary data has been collected from standard textbooks, journals, internet.

- **Research instruments used:**

  The research instruments used for the analysis of the primary data consists of

  - K- nearest neighbor(K-NN)
  - Logistic regression model
  - Naïve Bayes classifier.

- **Data Balancing method used:**

  ROSE (Random Over-Sampling Examples) method is used to balance the data

- **Sample size:**

  1813 sample customer data is used

- **Tools used for Visualization and Analysis:**

  - Power BI
  - R programming
  - MS – Excel
  - MS - Word

# CHAPTER 4

# DATA VISUALIZATION AND PROCESSING

## 4.1 Dataset Information:

The data set consist of 20 attributes of which 19 are predictor variables and “interest in solar purchase” which is a Binary variable used as the dependent variable.

Total of 1813 customers data is present in this data set.

## 4.2 Visualization of the Dataset:

### 4.2.1 Customer solar purchase interest:

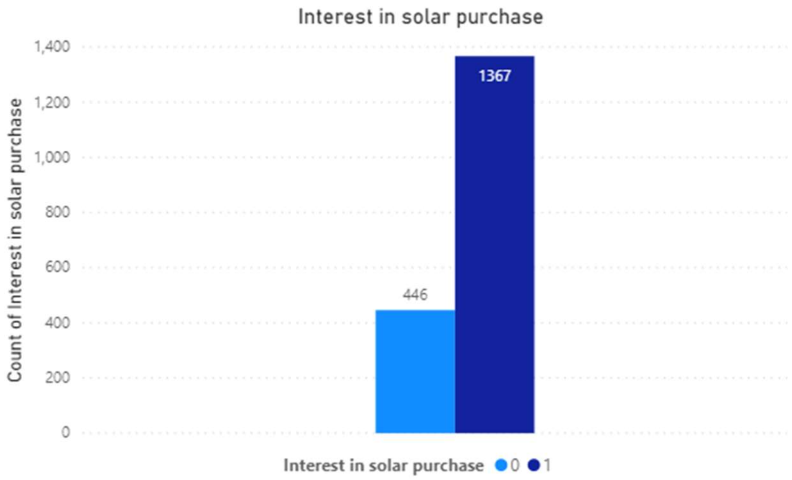


*Figure - 2*

The above graph shows the customer interest in purchasing solar panel, “0” represents Not Interested and “1” represents Interested.

### 4.2.2 Solar purchase interest by Age of customers:

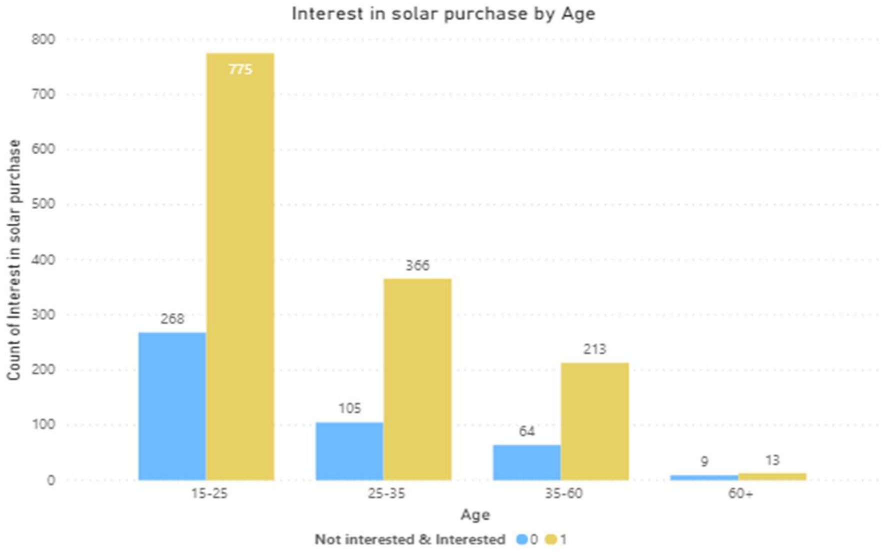


*Figure - 3*

- 74% of customers belonging to 15-25 age group has shown interest in solar panel purchase.
- 78% of customers belonging to 25-35 age group has shown interest in solar panel purchase
- 77% of customers belonging to 35-60 age group has shown interest in solar panel purchase
- 59% of customers belonging to 60+ age group has shown interest in solar panel purchase

### 4.2.3 Solar purchase interest based on the percentage of roof available for installation:

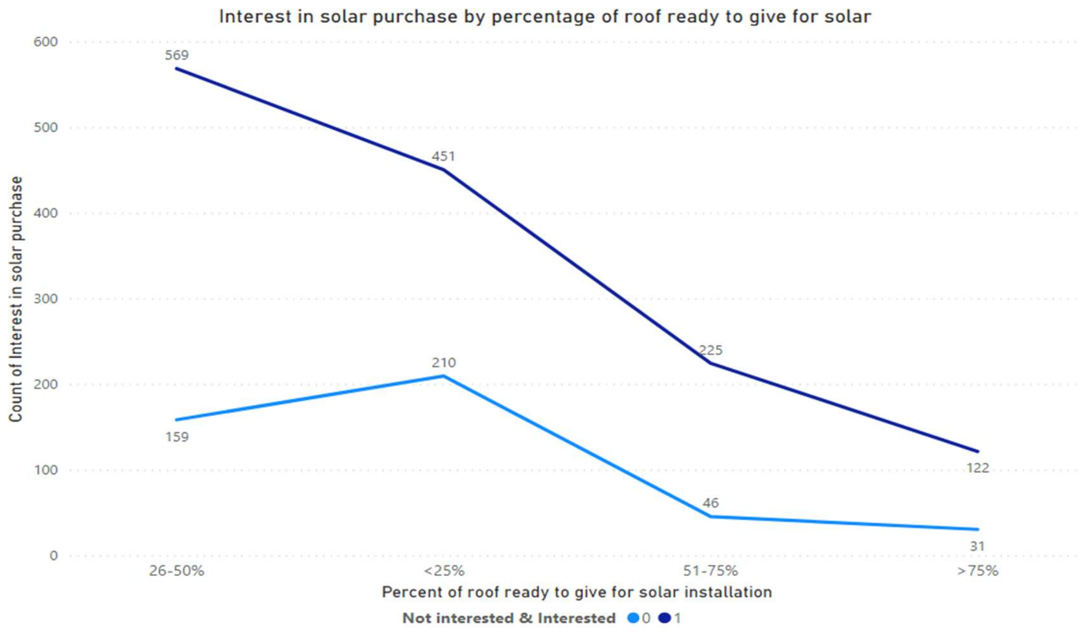


***Figure - 4***

The graph shows comparison between percentage of rooftop are available for solar installation in the customer house vs interest in solar system purchase.

Based on the graph

- 78% customers with 26-50% their home rooftop free are showing interest in solar system purchase
- 68% customer with <25% rooftop area free has shown interest in solar system purchase
- 83% customer with 51-75 % rooftop area free has shown interest in solar system purchase
- 80% customer with >75% rooftop is area has shown interest in solar system purchase

### 4.2.4 Solar purchase interest based on family average income:

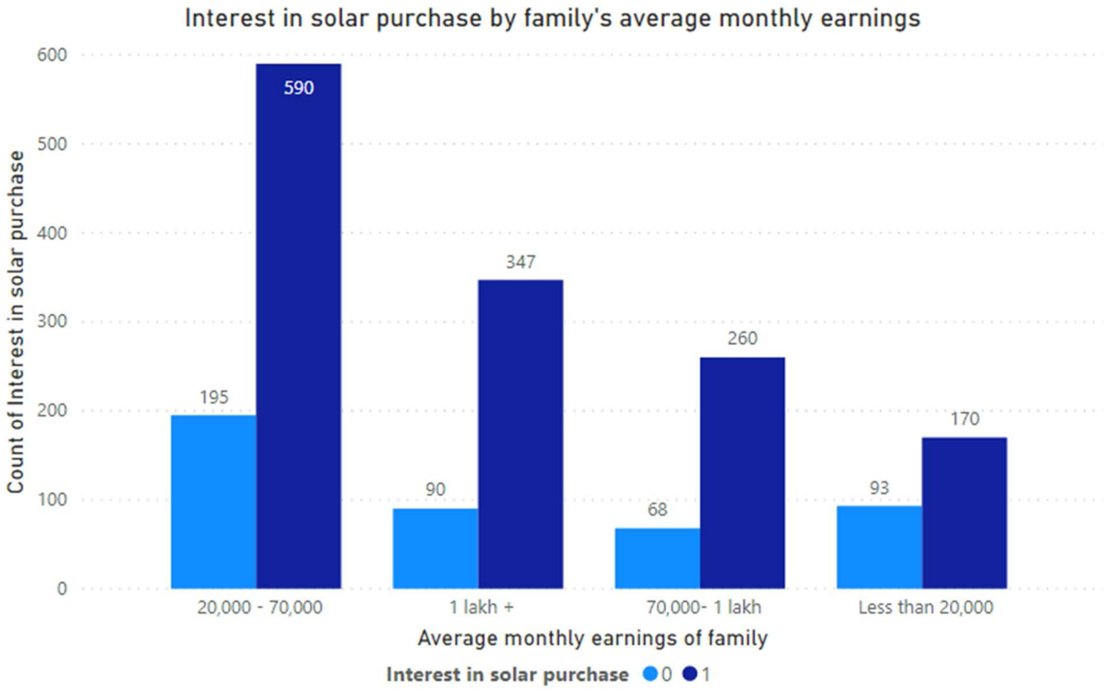


***Figure - 5***

- In the monthly income group 20k to 70k,75% people are interested in the home solar system purchase
- 79% people of the 1lakh+ income group are interest in purchase the solar panels.
- In the income group of 70k to 1 lakh,79% people are interest in purchasing solar panels.
- Of the less than 20k income group around 65% people are interested in the purchase of solar system for their homes.

### 4.2.5 Maximum investment interest to make in solar vs Count:

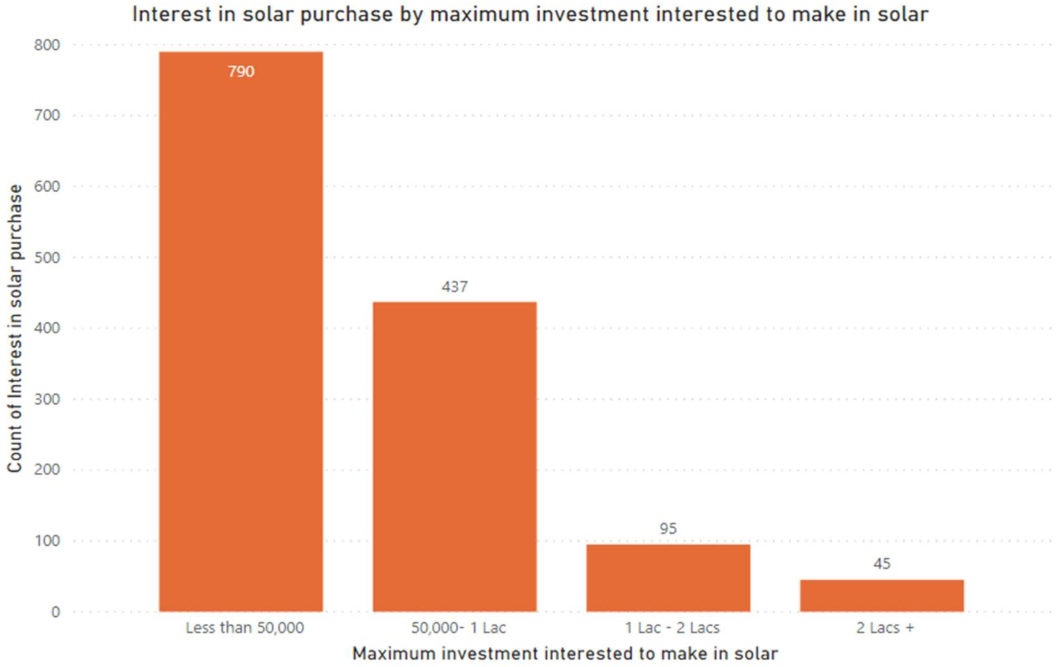


***Figure - 6***

- From the above graph we can interpret that high number of people, (790 of those interested in purchase) interested to make investment in solar system want to invest Rs. 50,000 or less.

- 437 people are interested to make an investment worth between Rs 50,000 to 1 Lac.

- Only 45 of the interested people want to make investments worth 2 Lac or more.

### 4.2.6 Purchase interest vs Level of awareness about solar:

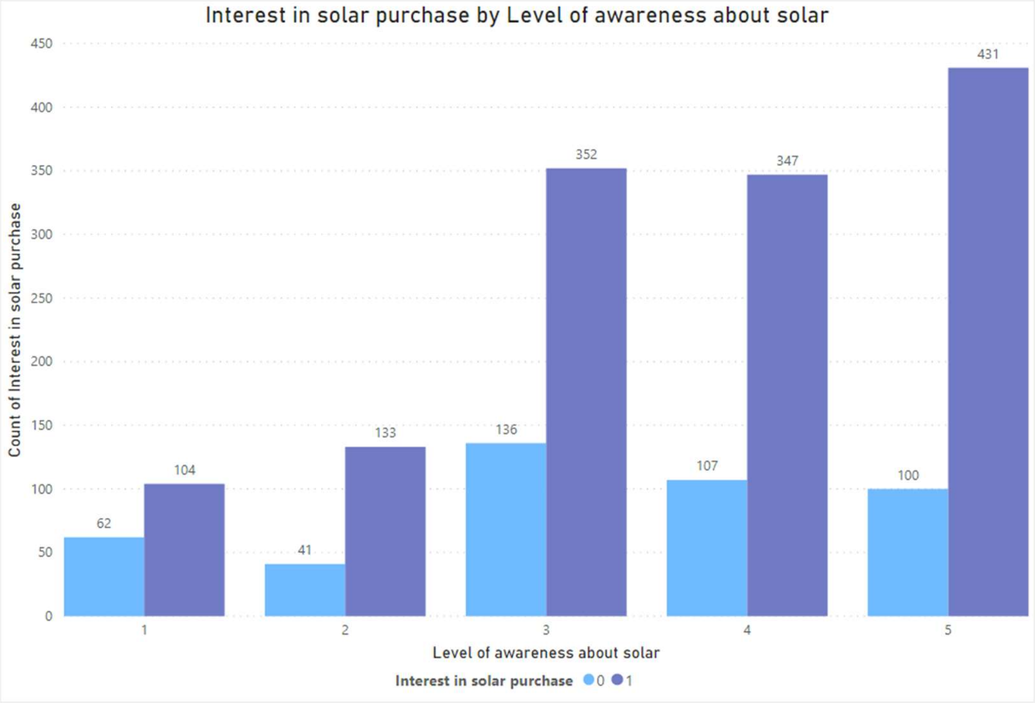


*Figure - 7*

The level of awareness is in increasing order with "1" being the least and "5" being the highest level of awareness .

- Based on the graph we can understand that the level of awareness does impact highly on a person's purchase interest of the solar system.
- Of the peoples with level 5 of awareness about solar, around 81% are interested in purchasing the solar system for their homes
- 76% of the people with level 4 awareness is interested in the purchase of the panels.
- 73% of the people with level 3awarness is interested in purchasing solar panels for their homes
- Around 62% people having awareness level of solar as 1 is interested in purchasing solar panels for their homes.

### 4.2.7 Lowest expected saving from electricity by installing solar panels Vs count:

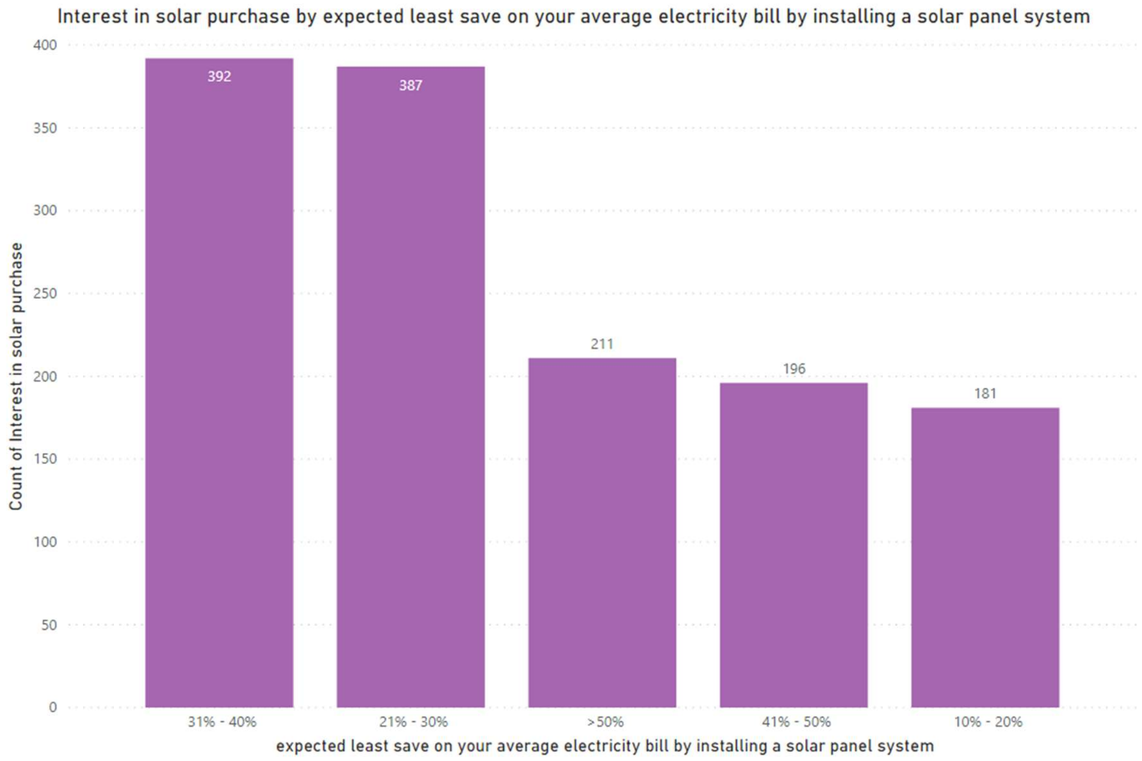


***Figure - 8***

- 392 of the Customers interested in purchasing solar panels expect 31% to 40% savings in their electricity bills.
- 387 of the Customers interested in purchasing solar panels expect 21% to 30% savings in their electricity bills.
- 211 of the Customers interested in purchasing solar panels expect greater than 50% savings in their electricity bills.
- 196 of the Customers interested in purchasing solar panels expect 41% to 50% savings in their electricity bills.
- 181 of the Customers interested in purchasing solar panels expect 10% to 200% savings in their electricity bills.

### 4.2.8 Purchase interest with respect to relative/friend who bought solar system:

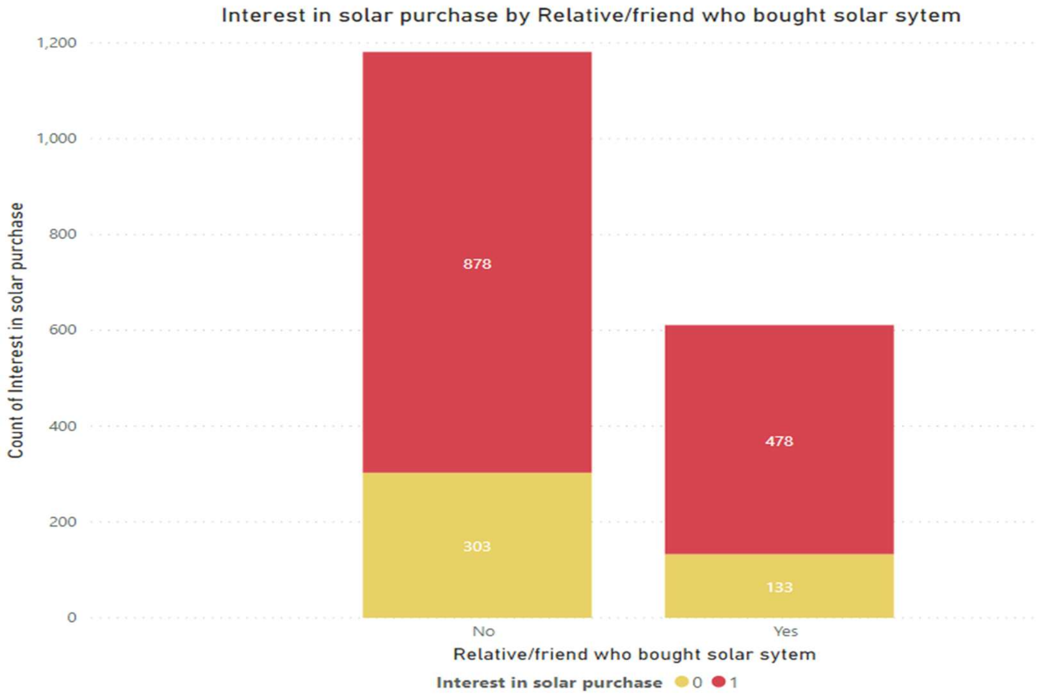


***Figure – 9***

The above graph shows how the purchase interest of a people is affected based on relative/friend who bought solar system**.**

- There seems to be less effect on the purchase interest based on relative/friend who bought solar system.
- 74% of the people with relatives/friends who doesn't buy solar system are still interested in purchasing solar systems for their homes
- 78% of the people with relatives/friends who brought solar system are interested in purchasing solar panels

### 4.3 Splitting Data for Model Training and Testing:

The entire data set is split in to training and testing sets, this is done to train the model on the training data set and test the model fit with testing data set.

The data is split in the ratio 70:30, ie 70% data into training set and remaining 30% data into testing set.

**Imbalanced Training Dataset:**

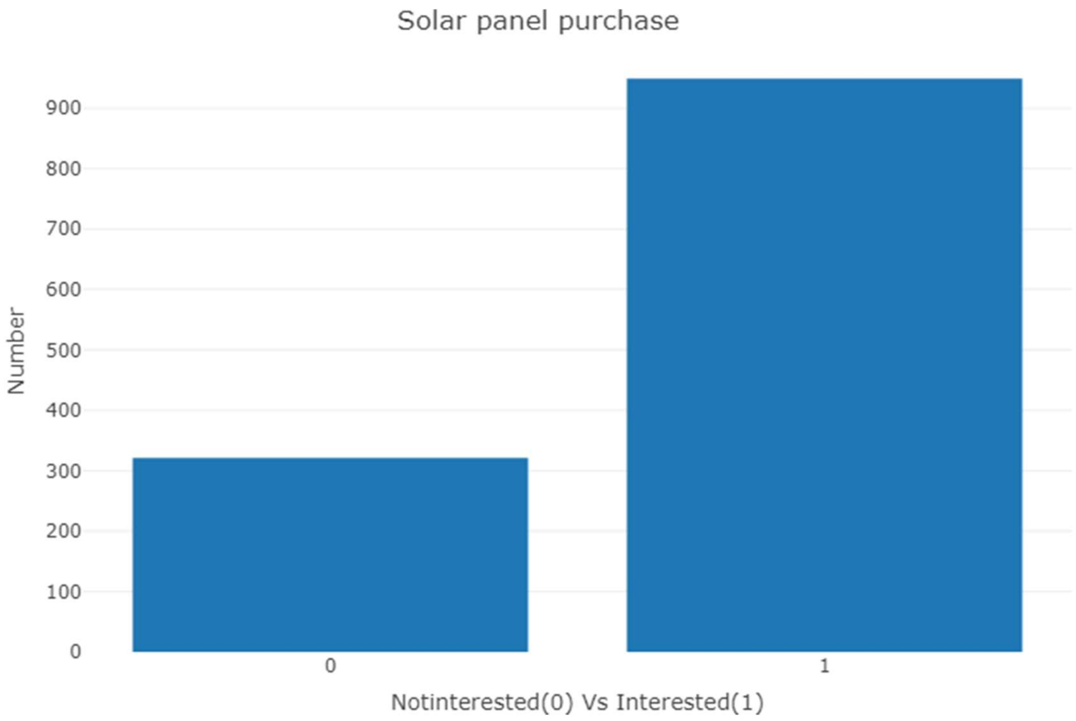


***Figure – 10***

- After splitting, the training data set consists of solar purchase interest as seen in the above graph.
- The data is highly imbalanced data, with 949 customers interested in purchasing solar system and 321 customers not interested in purchasing solar systems for their homes.
- The training data set needs to be balanced, ie both not interested and interested numbers showed be almost equal before apply in the prediction model. This is because prediction models like logistic regression, naïve bayes cannot handle imbalanced data and hence their prediction accuracy is highly affected.

### 4.4 Balancing of Training Data Set:

Balancing dataset can be done in many ways

1. Under Sampling
2. Over Sampling
3. Synthetic Data Generation

**Under Sampling Method**: This a typical sampling method mostly used when the data set is huge. It works by reducing the majority class data and balance it with minority class data.

This method removes original data hence is suitable only when we are dealing with data set with huge number of observations or rows.

**Over Sampling Method:** This method is also called Up sampling, it works with the minority class and replicates the minority class values till it balances with the majority class.The drawback with is method is the that it ends up adding same data multiple times to balance, hence the model accuracy is affected

**Synthetic Data Generation Method:** This is most commonly used data balancing techniques, in this method instead of adding minority class data multiple times or removing majority class data to balance the data set, it generates artificial data based on the metrices in the data and balances the dataset.

ROSE (Random Over-Sampling Examples) method is a type of synthetic data generation which used in this analysis to balance the training data set.

### 4.4.1 Balanced Training Data Set:

With the help of ROSE sampling method, the imbalanced training data set is now almost balanced with the addition of synthetic generated data.

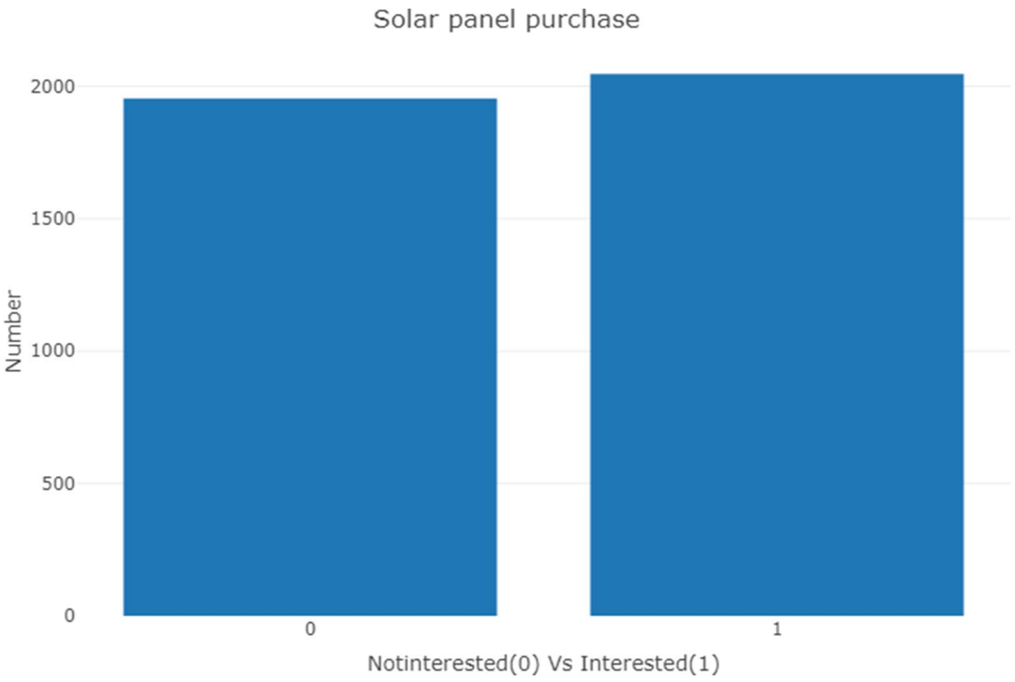
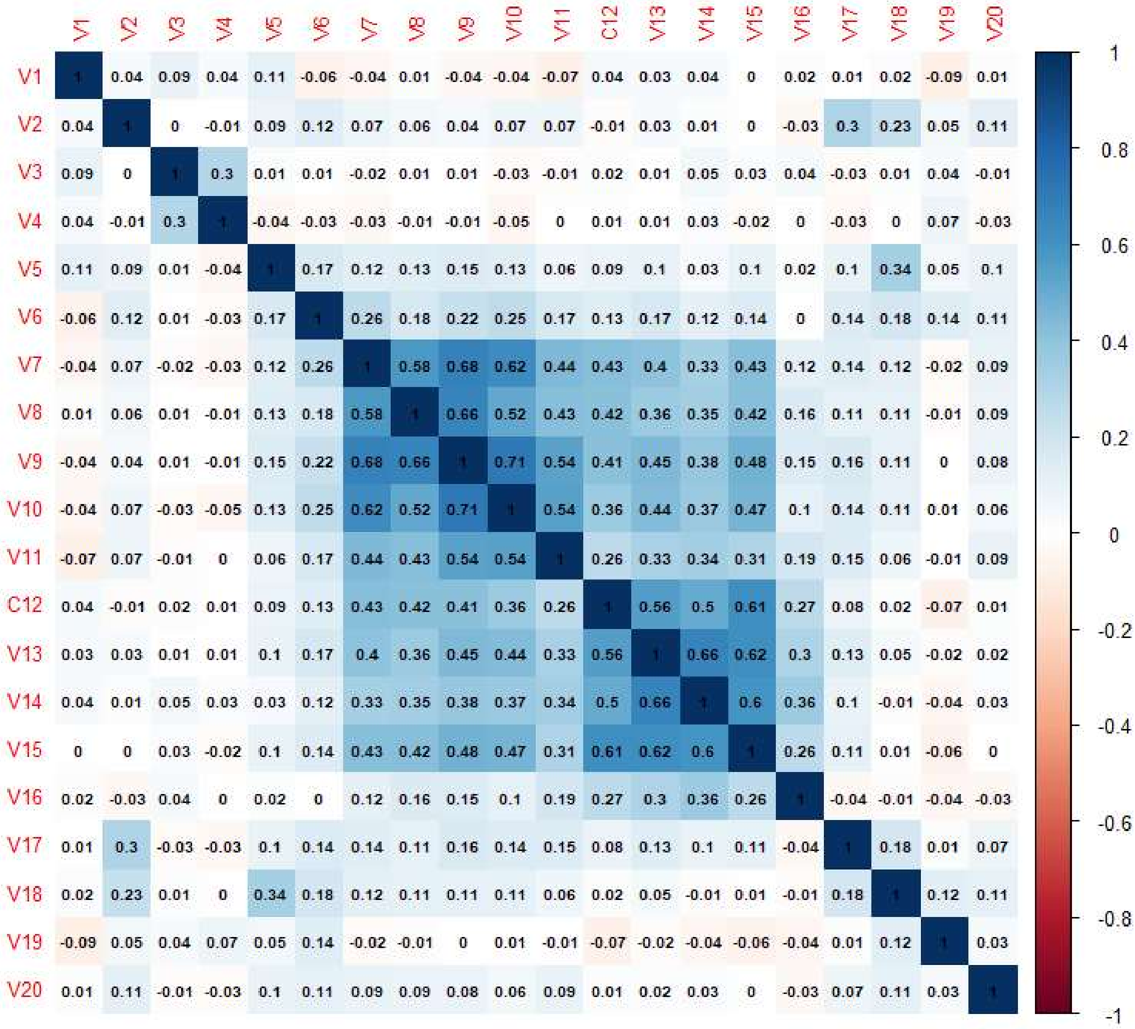


***Figure – 11***

- Due to the generation of synthetic data, the overall data size is also increased.
- Now the total number of observations in the training data set is 4000
- With 1954 observations of Not interested customers and 2046 observations of people who are interested in solar panel purchase.

### 4.4.2 Correlation Matrix of the Data Set:



| | V1 | V2 | V3 | V4 | V5 | V6 | V7 | V8 | V9 | V10 | V11 | C12 | V13 | V14 | V15 | V16 | V17 | V18 | V19 | V20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| V1 | 1 | 0.04 | 0.09 | 0.04 | 0.11 | -0.06 | -0.04 | 0.01 | -0.04 | -0.04 | -0.07 | 0.04 | 0.03 | 0.04 | 0 | 0.02 | 0.01 | 0.02 | -0.09 | 0.01 |
| V2 | 0.04 | 1 | 0 | -0.01 | 0.09 | 0.12 | 0.07 | 0.06 | 0.04 | 0.07 | 0.07 | -0.01 | 0.03 | 0.01 | 0 | -0.03 | 0.3 | 0.23 | 0.05 | 0.11 |
| V3 | 0.09 | 0 | 1 | 0.3 | 0.01 | 0.01 | -0.02 | 0.01 | 0.01 | -0.03 | -0.01 | 0.02 | 0.01 | 0.05 | 0.03 | 0.04 | -0.03 | 0.01 | 0.04 | -0.01 |
| V4 | 0.04 | -0.01 | 0.3 | 1 | -0.04 | -0.03 | -0.03 | -0.01 | -0.01 | -0.05 | 0 | 0.01 | 0.01 | 0.03 | -0.02 | 0 | -0.03 | 0 | 0.07 | -0.03 |
| V5 | 0.11 | 0.09 | 0.01 | -0.04 | 1 | 0.17 | 0.12 | 0.13 | 0.15 | 0.13 | 0.06 | 0.09 | 0.1 | 0.03 | 0.1 | 0.02 | 0.1 | 0.34 | 0.05 | 0.1 |
| V6 | -0.06 | 0.12 | 0.01 | -0.03 | 0.17 | 1 | 0.26 | 0.18 | 0.22 | 0.25 | 0.17 | 0.13 | 0.17 | 0.12 | 0.14 | 0 | 0.14 | 0.18 | 0.14 | 0.11 |
| V7 | -0.04 | 0.07 | -0.02 | -0.03 | 0.12 | 0.26 | 1 | 0.58 | 0.68 | 0.62 | 0.44 | 0.43 | 0.4 | 0.33 | 0.43 | 0.12 | 0.14 | 0.12 | -0.02 | 0.09 |
| V8 | 0.01 | 0.06 | 0.01 | -0.01 | 0.13 | 0.18 | 0.58 | 1 | 0.66 | 0.52 | 0.43 | 0.42 | 0.36 | 0.35 | 0.42 | 0.16 | 0.11 | 0.11 | -0.01 | 0.09 |
| V9 | -0.04 | 0.04 | 0.01 | -0.01 | 0.15 | 0.22 | 0.68 | 0.66 | 1 | 0.71 | 0.54 | 0.41 | 0.45 | 0.38 | 0.48 | 0.15 | 0.16 | 0.11 | 0 | 0.08 |
| V10 | -0.04 | 0.07 | -0.03 | -0.05 | 0.13 | 0.25 | 0.62 | 0.52 | 0.71 | 1 | 0.54 | 0.36 | 0.44 | 0.37 | 0.47 | 0.1 | 0.14 | 0.11 | 0.01 | 0.06 |
| V11 | -0.07 | 0.07 | -0.01 | 0 | 0.06 | 0.17 | 0.44 | 0.43 | 0.54 | 0.54 | 1 | 0.26 | 0.33 | 0.34 | 0.31 | 0.19 | 0.15 | 0.06 | -0.01 | 0.09 |
| C12 | 0.04 | -0.01 | 0.02 | 0.01 | 0.09 | 0.13 | 0.43 | 0.42 | 0.41 | 0.36 | 0.26 | 1 | 0.56 | 0.5 | 0.61 | 0.27 | 0.08 | 0.02 | -0.07 | 0.01 |
| V13 | 0.03 | 0.03 | 0.01 | 0.01 | 0.1 | 0.17 | 0.4 | 0.36 | 0.45 | 0.44 | 0.33 | 0.56 | 1 | 0.66 | 0.62 | 0.3 | 0.13 | 0.05 | -0.02 | 0.02 |
| V14 | 0.04 | 0.01 | 0.05 | 0.03 | 0.03 | 0.12 | 0.33 | 0.35 | 0.38 | 0.37 | 0.34 | 0.5 | 0.66 | 1 | 0.6 | 0.36 | 0.1 | -0.01 | -0.04 | 0.03 |
| V15 | 0 | 0 | 0.03 | -0.02 | 0.1 | 0.14 | 0.43 | 0.42 | 0.48 | 0.47 | 0.31 | 0.61 | 0.62 | 0.6 | 1 | 0.26 | 0.11 | 0.01 | -0.06 | 0 |
| V16 | 0.02 | -0.03 | 0.04 | 0 | 0.02 | 0 | 0.12 | 0.16 | 0.15 | 0.1 | 0.19 | 0.27 | 0.3 | 0.36 | 0.26 | 1 | -0.04 | -0.01 | -0.04 | -0.03 |
| V17 | 0.01 | 0.3 | -0.03 | -0.03 | 0.1 | 0.14 | 0.14 | 0.11 | 0.16 | 0.14 | 0.15 | 0.08 | 0.13 | 0.1 | 0.11 | -0.04 | 1 | 0.18 | 0.01 | 0.07 |
| V18 | 0.02 | 0.23 | 0.01 | 0 | 0.34 | 0.18 | 0.12 | 0.11 | 0.11 | 0.11 | 0.06 | 0.02 | 0.05 | -0.01 | 0.01 | -0.01 | 0.18 | 1 | 0.12 | 0.11 |
| V19 | -0.09 | 0.05 | 0.04 | 0.07 | 0.05 | 0.14 | -0.02 | -0.01 | 0 | 0.01 | -0.01 | -0.07 | -0.02 | -0.04 | -0.06 | -0.04 | 0.01 | 0.12 | 1 | 0.03 |
| V20 | 0.01 | 0.11 | -0.01 | -0.03 | 0.1 | 0.11 | 0.09 | 0.09 | 0.08 | 0.06 | 0.09 | 0.01 | 0.02 | 0.03 | 0 | -0.03 | 0.07 | 0.11 | 0.03 | 1 |

*Figure – 12*

The above plot shows the correlation of all the 20 attributes in the data set.

- **V1 -** age
- **V2 -** percentage of roof ready to give for solar
- **V3 -** Do you have a home loan
- **V4 -** currently have an EMI for any home appliance
- **V5 -** family's average monthly earnings range
- **V6 -** How much aware are you about solar?
- **V7 -** importance in your solar purchase decision-[Power backup]
- **V8 -** importance in your solar purchase decision – [electricity prices]

- **V9 -** importance in your solar purchase decision – [good saving on the power bill]
- **V10 -** importance in your solar purchase decision - [Environment friendly]
- **V11 -** importance in your solar purchase decision- [Earn from empty roof]
- **V12 -** To what extent the factors discourage you from installing solar- [years it takes for the investment to pay back]
- **V13 -** To what extent factors discourage you from installing solar- [Lack of government incentives]
- **V14 -** To what extent factors discourage you from installing solar- [Lack of appropriate loan options]
- **V15 -** To what extent factors discourage you from installing solar- [High cost of solar power systems]
- **V16 -** To what extent factors discourage you from installing solar- [My rooftop is not suitable for solar]
- **V17 -** expected least save on your average electricity bill by installing a solar panel system
- **V18 -** What maximum investment would you be willing to make in solar?
- **V19 -** Do you have any relative/friend that bought a solar system?
- **V20 -** Interest in solar purchase

**Interpretation of Correlation Matrix:**

- The white or near white color represents no to very low correlation between the attributes
- Light blue to dark blue represents moderate to high positive correlation among the two attributes
- Light red to Dark red represents Moderated to High negative correlation among two attributes.
- The numerical value in the color boxes represents the correlation coefficients. The values range between -1 and +1, where 0 to -1 indicates negative correlation and 0 to +1 indicates positive correlation.

# CHAPTER 5

# ANALYSIS AND INTERPRETATION

After training the model with the balanced training dataset, the previously split testing data set is used to analysis the model accuracy and fit.

**LOGISTIC REGRESSION:**

It is a type of classification model used when the dependent variable is binary like true/false,0/1, Yes/No. Logistic regression tries to fit the model in a sigmoid curve which is "S" shaped curve

$$\mathrm{p} = \frac{1}{1+ e^{-y}}$$

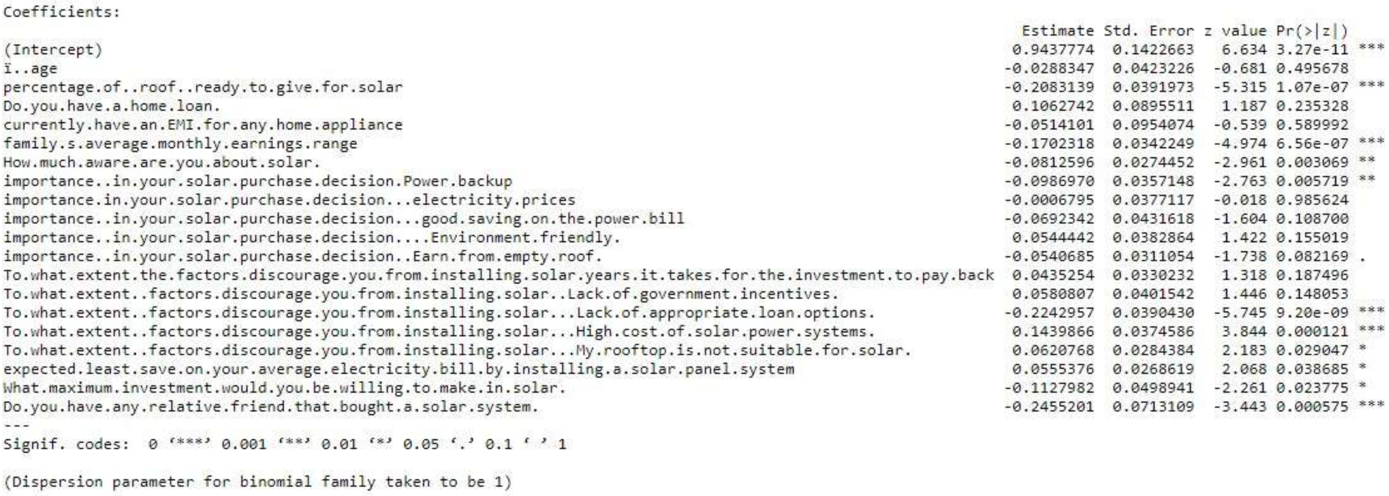
The logistic regression finds the probability of an even occurring or not occurring.

y = a+ bx (linear regression equation), 'a' being the intercept and 'b' being the slope of the line.

Final equation when linked with linear regression equation:

$$\ln\left(\frac{p}{1-p}\right) = a + bx$$

### 5.1 Logistic Regression Model:



```
Coefficients:
                                                                                                                Estimate Std. Error z value Pr(>|z|)    
(Intercept)                                                                                                    0.9437774  0.1422663   6.634 3.27e-11 ***
ï..age                                                                                                        -0.0288347  0.0423226  -0.681 0.495678    
percentage.of..roof..ready.to.give.for.solar                                                                  -0.2083139  0.0391973  -5.315 1.07e-07 ***
Do.you.have.a.home.loan.                                                                                       0.1062742  0.0895511   1.187 0.235328    
currently.have.an.EMI.for.any.home.appliance                                                                  -0.0514101  0.0954074  -0.539 0.589992    
family.s.average.monthly.earnings.range                                                                       -0.1702318  0.0342249  -4.974 6.56e-07 ***
How.much.aware.are.you.about.solar.                                                                           -0.0812596  0.0274452  -2.961 0.003069 ** 
importance..in.your.solar.purchase.decision.Power.backup                                                      -0.0986970  0.0357148  -2.763 0.005719 ** 
importance.in.your.solar.purchase.decision...electricity.prices                                               -0.0006795  0.0377117  -0.018 0.985624    
importance..in.your.solar.purchase.decision...good.saving.on.the.power.bill                                   -0.0692342  0.0431618  -1.604 0.108700    
importance..in.your.solar.purchase.decision....Environment.friendly.                                           0.0544442  0.0382864   1.422 0.155019    
importance..in.your.solar.purchase.decision..Earn.from.empty.roof.                                            -0.0540685  0.0311054  -1.738 0.082169 .  
To.what.extent.the.factors.discourage.you.from.installing.solar.years.it.takes.for.the.investment.to.pay.back  0.0435254  0.0330232   1.318 0.187496    
To.what.extent..factors.discourage.you.from.installing.solar..Lack.of.government.incentives.                   0.0580807  0.0401542   1.446 0.148053    
To.what.extent..factors.discourage.you.from.installing.solar...Lack.of.appropriate.loan.options.              -0.2242957  0.0390430  -5.745 9.20e-09 ***
To.what.extent..factors.discourage.you.from.installing.solar...High.cost.of.solar.power.systems.               0.1439866  0.0374586   3.844 0.000121 ***
To.what.extent..factors.discourage.you.from.installing.solar...My.rooftop.is.not.suitable.for.solar.           0.0620768  0.0284384   2.183 0.029047 *  
expected.least.save.on.your.average.electricity.bill.by.installing.a.solar.panel.system                        0.0555376  0.0268619   2.068 0.038685 *  
What.maximum.investment.would.you.be.willing.to.make.in.solar.                                                -0.1127982  0.0498941  -2.261 0.023775 *  
Do.you.have.any.relative.friend.that.bought.a.solar.system.                                                   -0.2455201  0.0713109  -3.443 0.000575 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

***Table – 1***

- The Logistic regression model is run with Target variable "Interest in solar purchase" which is a binary variable with 0 implying not interested and 1 implying interested in purchase.
- All the 19 predictor variables are used initially to compute the model.
- The above table after computing show the significance of each predictor variable, which is marked with stars and dots. More number of stars represents less significance, dots and blank represent high significance.

- The predictor variable with least significance is used to cross validate the regression model to optimize the accuracy of prediction.
- After computing with different combinations of less significant variables, the predictor variables which together is giving higher accuracy is noted below:
    - percentage of roof ready to give for solar
    - family's average monthly earnings range
    - How much aware are you about solar
    - To what extent factors discourage you from installing solar- [Lack of appropriate loan options]
    - Do you have any relative/friend that bought a solar system?

After cross validation using the above-mentioned variables, the confusion matrix and result generated is given below.

### 5.1.1 Logistic Regression Confusion Matrix After Cross Validation:

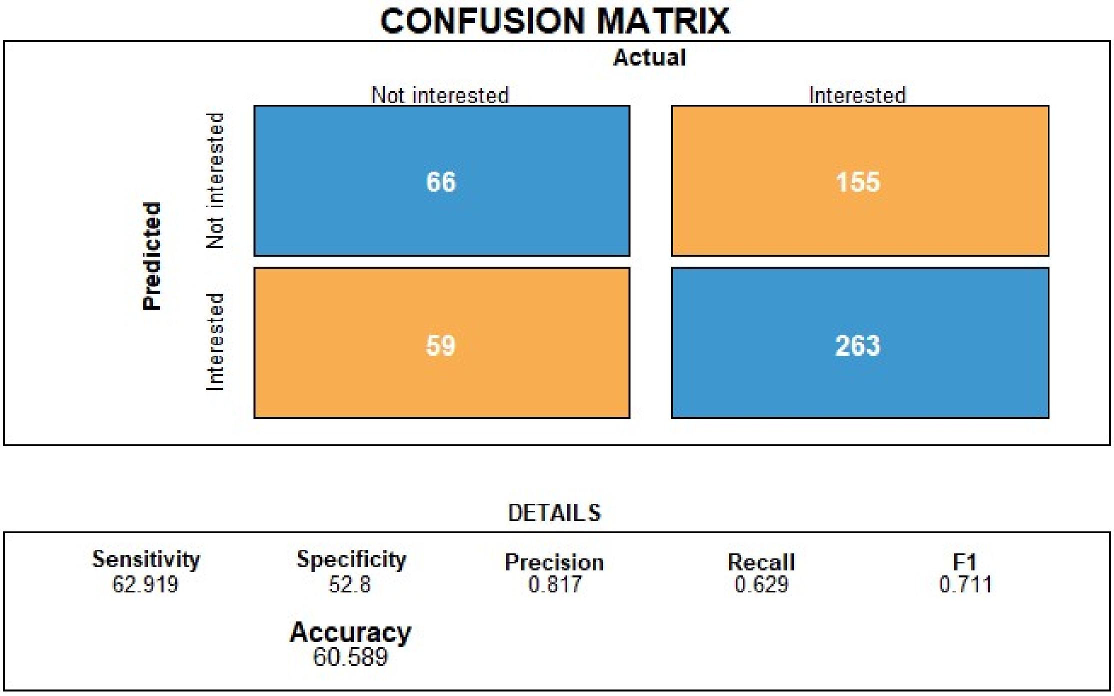


*Figure -13*

Positive class in the confusion matrix is "Interested"

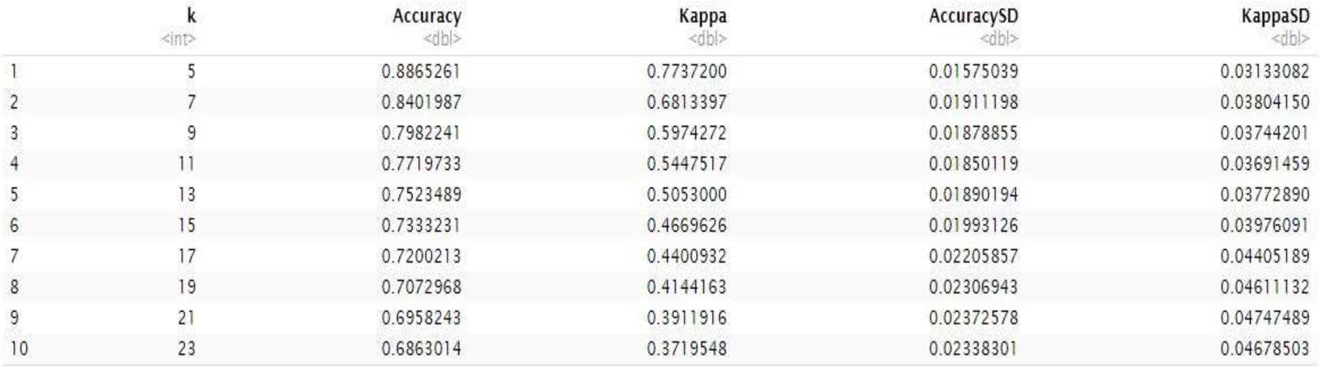
The confusion matrix shows four types of values:

- TP (True Positive): predicted customer is interested in solar panel purchase and actually the customer is interested in purchase
- TN (True Negative): predicted customer is not interested in solar system purchase
- FP (False Positive): Predicted customer is interested, but actually not interested
- FN (False Negative): Predicted customer is not interested, but actually interested

- ✓ Precision
  precision = (TP) / (TP+FP)
- ✓ Recall/ Sensitivity
  Recall/ sensitivity = (TP) / (TP+FN)
- ✓ Accuracy
  accuracy = (TP + TN) / (TP + TN + FP + FN)
- ✓ Specificity
  Specificity = TN / (TN + FP)

Logistic regression model is having 60% accuracy, means that the model is 60% accurate in identifying whether customer is interested in solar purchase or Not.
The model has sensitivity of 62.9%, This shows that roughly 63% of the data is correctly classified interested.
The model is having 52.8% specificity, this means roughly 47% customers who are actually not interested in solar system purchase are incorrectly predicted as Interested.

### 5.2 K-Nearest Neighbor (K-NN):



|  | k<br><int> | Accuracy<br><dbl> | Kappa<br><dbl> | AccuracySD<br><dbl> | KappaSD<br><dbl> |
|---|---|---|---|---|---|
| 1 | 5 | 0.8865261 | 0.7737200 | 0.01575039 | 0.03133082 |
| 2 | 7 | 0.8401987 | 0.6813397 | 0.01911198 | 0.03804150 |
| 3 | 9 | 0.7982241 | 0.5974272 | 0.01878855 | 0.03744201 |
| 4 | 11 | 0.7719733 | 0.5447517 | 0.01850119 | 0.03691459 |
| 5 | 13 | 0.7523489 | 0.5053000 | 0.01890194 | 0.03772890 |
| 6 | 15 | 0.7333231 | 0.4669626 | 0.01993126 | 0.03976091 |
| 7 | 17 | 0.7200213 | 0.4400932 | 0.02205857 | 0.04405189 |
| 8 | 19 | 0.7072968 | 0.4144163 | 0.02306943 | 0.04611132 |
| 9 | 21 | 0.6958243 | 0.3911916 | 0.02372578 | 0.04747489 |
| 10 | 23 | 0.6863014 | 0.3719548 | 0.02338301 | 0.04678503 |

***Table -2***

**Plot of K value:**

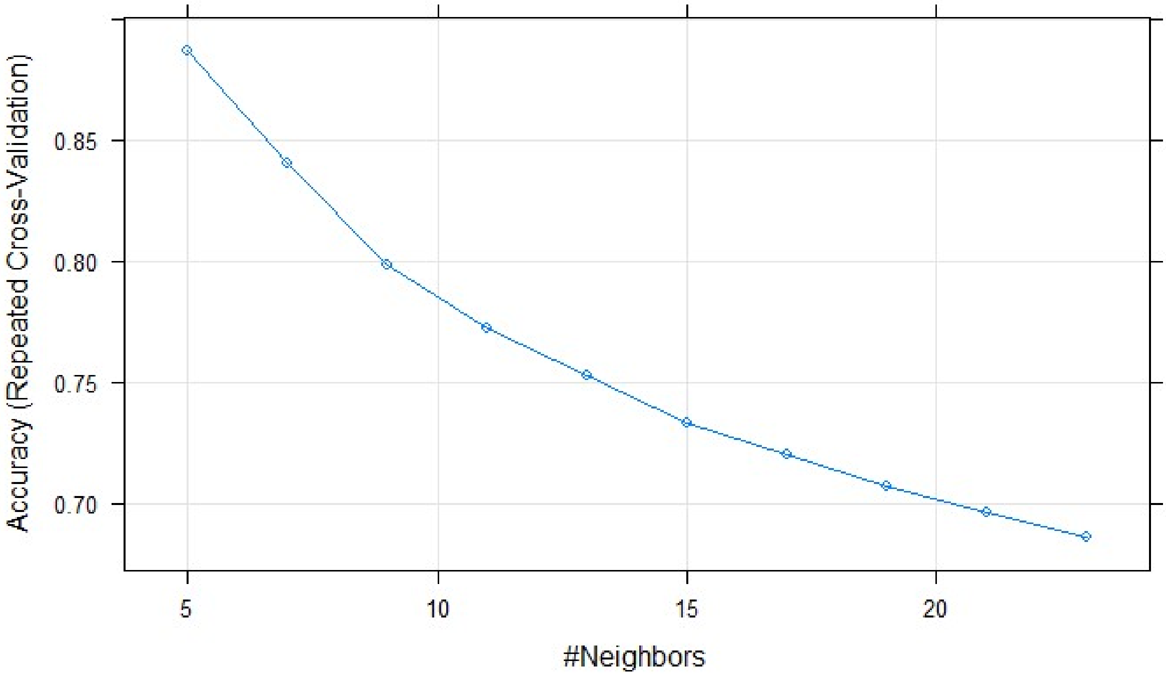
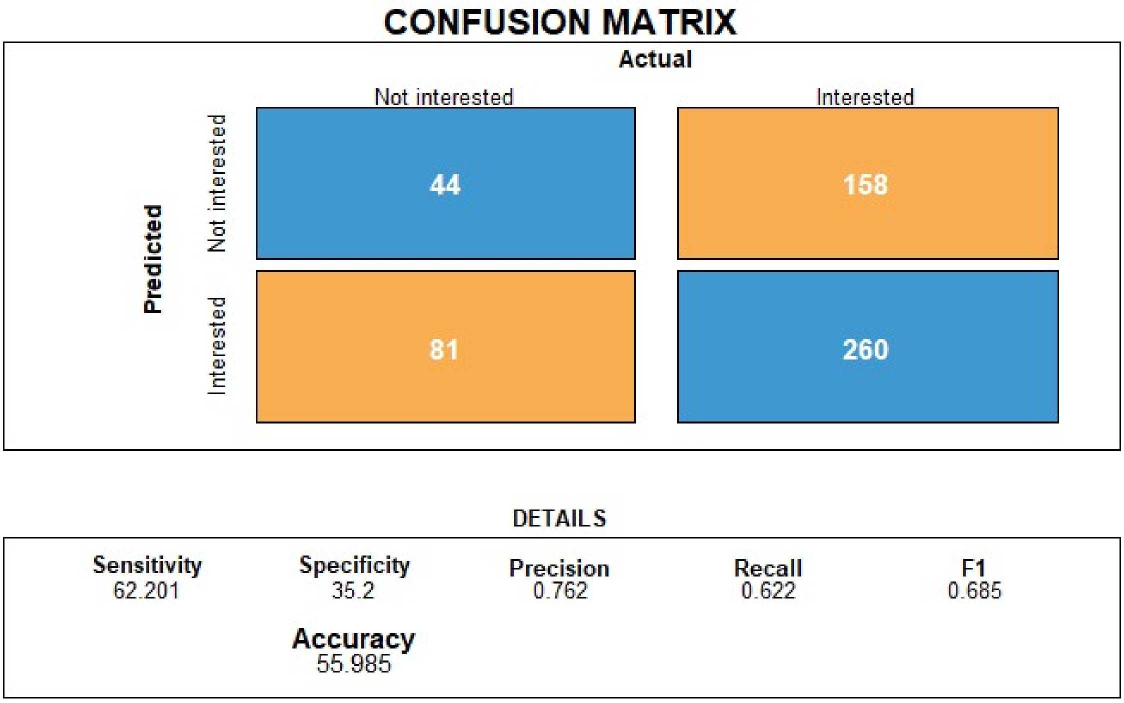


***Figure – 14***

After computing cross validation, the K value for the K-NN model is identified as above, K- value with highest accuracy is used in the model., ie K = 5

### 5.2.1 K-NN Confusion Matrix After Cross Validation:


CONFUSION MATRIX

| Predicted \ Actual | Not interested | Interested |
|---|---|---|
| Not interested | 44 | 158 |
| Interested | 81 | 260 |

DETAILS

| Sensitivity | Specificity | Precision | Recall | F1 |
|---|---|---|---|---|
| 62.201 | 35.2 | 0.762 | 0.622 | 0.685 |

Accuracy
55.985

***Figure – 15***

Positive class in the confusion matrix is “Interested”

The K-NN model has generated accuracy of 55.9% after cross validation

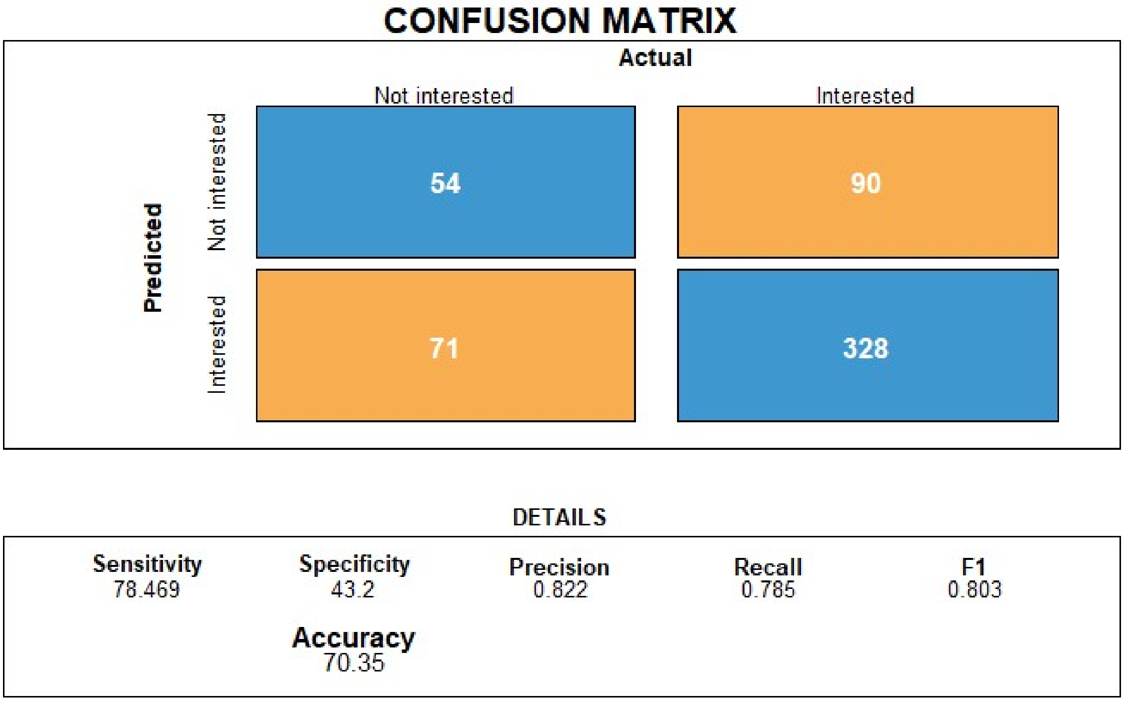
The model has sensitivity of 62.2%, This shows that roughly 62% of the data is correctly classified interested.

The model is having 35.2% specificity.

### 5.3 Naïve Bayes Model:

After cross validation the confusion matrix is generated.

#### 5.3.1 Naive Bayes Confusion Matrix After cross validation:



CONFUSION MATRIX

| Predicted \ Actual | Not interested | Interested |
|---|---|---|
| Not interested | 54 | 90 |
| Interested | 71 | 328 |

DETAILS

| Sensitivity | Specificity | Precision | Recall | F1 |
|---|---|---|---|---|
| 78.469 | 43.2 | 0.822 | 0.785 | 0.803 |

Accuracy
70.35

*Figure – 16*

Positive class in the confusion matrix is “Interested”

The Naïve bayes model has generated accuracy of 70.35%

The model has sensitivity of 78.46 %, This shows that roughly 79% of the data is correctly classified interested.

The model is having 43.2 % specificity, this means roughly 57 % customers who are actually not interested in solar panel system purchase are incorrectly predicted as Interested

# CHAPTER – 6
# CONCLUSION

The purpose of the creation of the prediction model was to help sales employees of the Peacock solar company to easily identify customers who are interested in the purchase of the solar panel system for their homes by providing the prediction model with the collected independent variables data that is mentioned in the (table-1) from the customers. With the help of these data collected from the customer the model will be able to predict if the customer is likely to be purchase the solar panel.

The prediction models used were Logistic regression model, K-NN and Naïve Bayes classifier models. Of the three models Naïve bayes model seems to be performing better with 70% accuracy and 78% sensitivity in identifying interested customer class. Logistic regression model is also good but, the accuracy rated is considerably lower than Naïve Bayes model.

For the model to perform better, more numbers of data are needed to train the model. So over time training the model with more amounts of data helps in improving the prediction accuracy. Using both the logistic regression and naïve bayes model to interpret the customer interest would be better.

# CONFIDENTIALITY

All information gathered from this study was kept confidential. Participant's identity was not disclosed outside. The result of this study may be published for scientific purposes, and participant's identity will not be revealed.

# REFERENCES

- Eunju Kim, Wooju Kim, & Yillbyung Lee (2002) “Combination of multiple classifiers for the customer’s purchase behavior prediction”, Decision Support Systems 34 (2002) 167–175.
- Chelsea Schelly (2014) “Residential solar electricity adoption: What motivates, and what matters? A case study of early adopters”, Energy Research & Social Science 2 (2014) 183–191.
- Serhat Peker, Altan Kocyigit and P. Erhan Eren (2017), “A hybrid approach for predicting customers’ individual purchase behavior” Kybernetes Vol. 46 No. 10, (2017) pp. 1614-1631
- C.Robert Newberry,Bruce R. Klemz, & Christo Boshoff(2003), “Managerial Implication of predicting purchase behaviour from purchase intentions.A retail patronage case study,Journal of service marketing Vol:17 No. 6 2003, pp 609-620
- P. Anitha & Malini M. Patil (2019), “RFM model for customer purchase behavior using K-Means algorithm”, Journal of King Saud University - Computer and Information Sciences
- Haifeng Zhang, Yevgeniy Vorobeychik, Joshua Letchford and Kiran Lakkaraju(2014), “Predicting Rooftop Solar Adoption Using Agent-Based Modeling” Energy Market Prediction: Papers from the 2014 AAAI Fall Symposium
- R Data Analysis Cookbook, by Viswa Viswanathan, Shanthi Viswanathan
- https://peacock.solar/