Bike Buyer Prediction

# Introduction:

Do you have any experience as a sales person in any exclusive show room? If not assume that as you are a sales person in any exclusive show room and you are dealing with a customer who had visited the store is who can’t afford it and you had tried your level best for making that customer to buy a product and he refuses to buy it. Then how do you feel after making that much effort and if it goes worthless?

If it happens same in the Motorcycle show room then applying the emerging technologies like machine learning is helpful? Here the motorcycle show room manager wants to know whether the customer buys a bike or not. For knowing that he can’t go near to a customer and can’t ask whether they are going to buy a bike or not? Instead of that by using the machine learning techniques just by asking some information of a customer like salary, no. of children and age etc he can predict that is he going to buy a bike or not?

## PYTHON:

Python is a high level programming language, which allows an programmer to focus on core functionality of the application by taking care of common programming tasks. The simple syntax rules of the programming language further makes it easier for you to keep the code base readable and application maintainable. Main reasons to use python language is:

1. Readable and Maintainable Code

2. Multiple Programming Paradigms

3. Compatible with Major Platforms and Systems

4.  Robust Standard Library

5. Many Open Source Frameworks and Tools

6. Simplify Complex Software Development

7. Adopt Test Driven Development.

## Machine learning

Machine learning is an subset of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

Basically Machine Learning is of three types:

1. Supervised Learning: Learning from characterized data.

2. Unsupervised Learning: Learning from raw data.

3. Reinforcement Learning: Learning from self mistakes (or) self learning data.

Predictive modeling is the way of building a model that is capable of making predictions. The process includes a machine learning algorithm that learns certain properties from a training dataset in order to make those predictions. Predictive modeling can be divided further into two areas: Regression and pattern classification. Regression models are based on the analysis of relationships between variables and trends in order to make predictions about continuous variables. In contrast to regression models ,the task of pattern classification is to assign discrete class labels to particular data value as output of a prediction Here we used Random Forest Classifier algorithm to predict whether the customer buys a motorcycle or not.

## Steps involved in Machine learning

Steps involved in the machine learning is shown in the below figure

1. Data collection

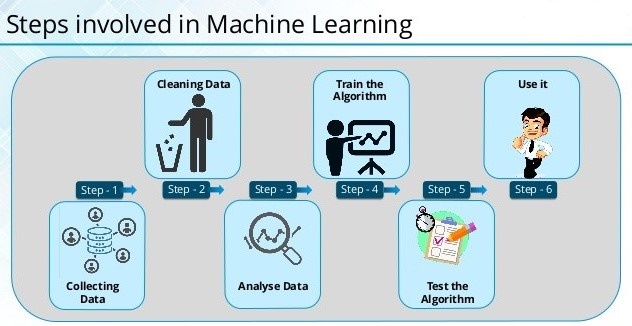
2. Data wrangling

3. Analyse the data

4. Train the algorithm

5. Test algorithm

6. Deployment



Machine learning is of mainly two models. They are Regression model and Classification model.

Here we are going to use the classification model since predicting whether a customer buys a bike or not produces the output as yes or no. The types of classification algorithms are:

1. Logistic Regression: Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable.
2. KNN model: K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). The k value must be always an odd number.
3. SVM model: “Support Vector Machine” (SVM) is a supervised machine learning algorithm . In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well. Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line).
4. Decision Tree Classifier: Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.
5. Random Forest Classifier: Random forest algorithm is a supervised classification algorithm. As the name suggest, this algorithm creates the forest with a number of trees. In general, the more trees in the forest the more robust the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives the high accuracy results.

# Objective of Research

The above problem made me to go for a research about how can we reduce the efforts of a sales person and managing time and to treat the customers for making easier his work. Through many documentation and cases, it came out that machine learning and data science can make the work easier and faster. The objective of this project is to analyze dataset which consist of the basic details of a customer who purchased a bike or not for the previous period.

# Problem Statement

To find out whether a customer is going to purchase a bike or not by using a classification model.

# 1.4. Industry Profile

As our project title is “Bike Buyer Prediction”, which is related to Motorcycle Show room where the customer profile like age, gender, income etc. are considered for predicting whether a customer is going to buy a bike or not?

# 2. Review of Literature

The dataset is taken from the following websites:

<https://studio.azureml.net/Home>

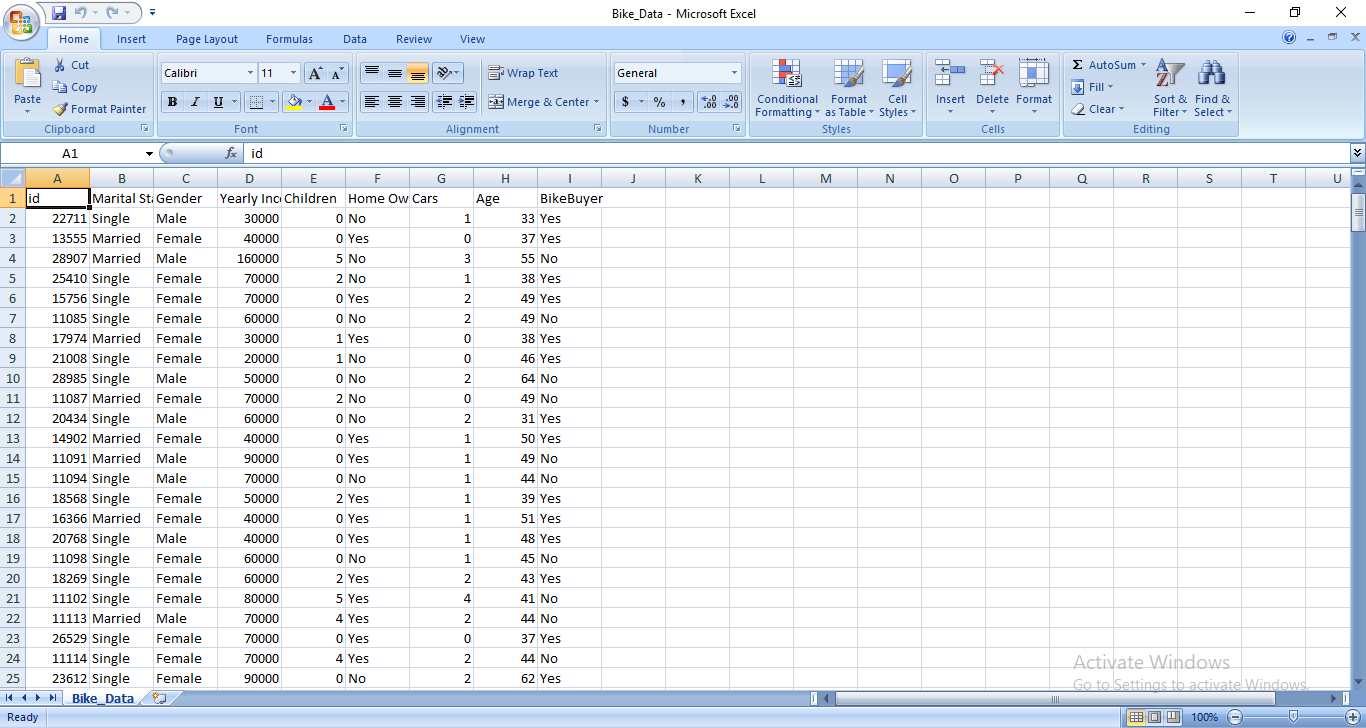
which is perfectly fitted to our classification model to predict whether a customer is going to buy a bike or not?

# 3. Data Collection

The data file format is in .csv form .The data fields which are used to predict health insurance are :

* Age
* Gender
* Children
* Income
* Home owner
* Cars owned
* Bike buyer
* There are two variables x and y where x is an independent variable which acts as an input where all the fields are taken.
* Where y is an dependent variable which has the last field i.e., Bike buyer which is predicted from independent variable x.
* The fields which comes under independent variables are:
* Age
* Gender
* Children
* Income
* Home owner
* Cars owned

All these independent variables acts as inputs to predict the Bike buyer.



# 4. Methodology

## 4.1. Exploratory Data Analysis

### Data collection:

Here we collect the data from the different motorcycle showrooms and we use that data and predict whether the customer is going to buy a bike or not.

### Data Preprocessing:

In this phase, the data is prepared for the analysis purpose which contains relevant information. Pre-processing and cleaning of data are one of the most important tasks that must be done one before dataset can be used for machine learning. The real-world data is noisy, incomplete and inconsistent. So, it is required to be cleaned.

### Extraction of Feature Set/Training Data

The feature set/ training data is extracted in the proportions of either 70% or 80%. For splitting the data we use the predefined methods in python.

### Implementation of Machine Learning Algorithm on Feature Set/Training Data

#### Regression:

A Regression model is created when we want to find out a number i.e., if the dependent variable is consists of continuous values. For example if we want to predict the cost of a used motorcycle since the cost varies from one to one we use the regression model for predicting.

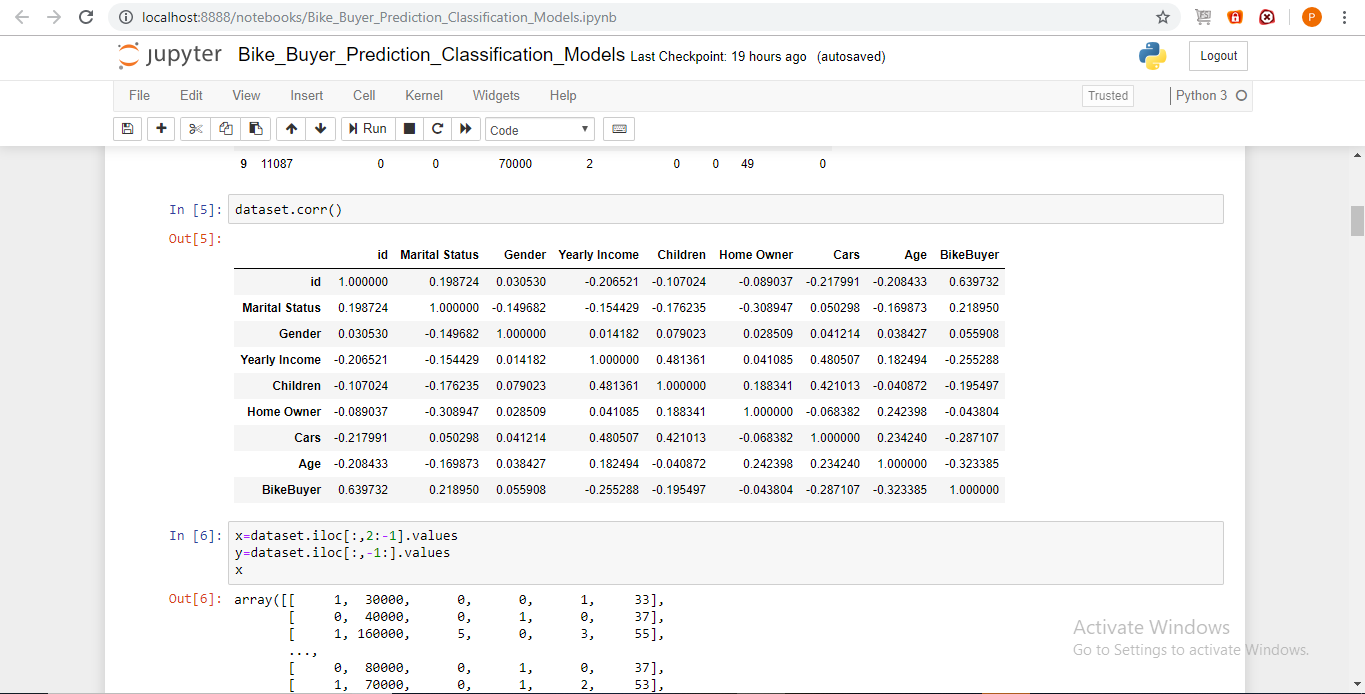
#### Classification:

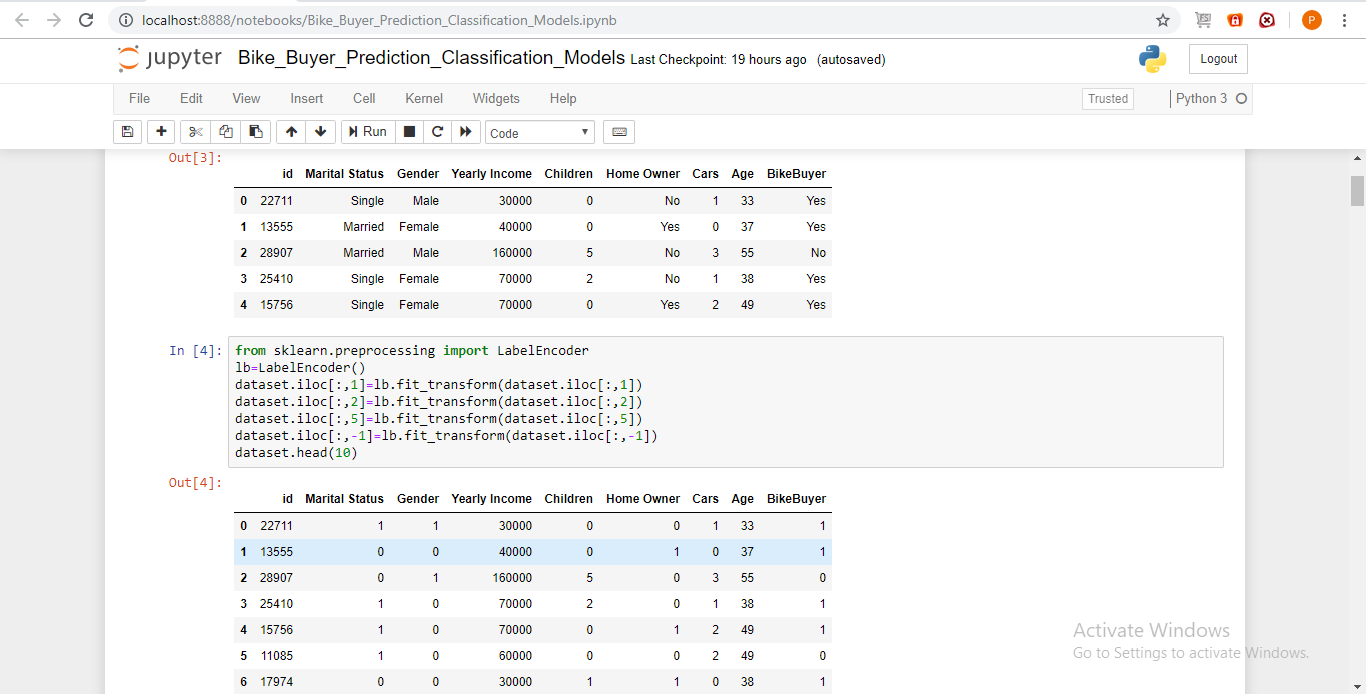
A classification model is created when the data is classified i.e., the dependent variable is classified into binary form or true/false or yes/no or fit/unfit etc. For example if we want to predict whether the customer is eligible for the loan.

### Testing of Data

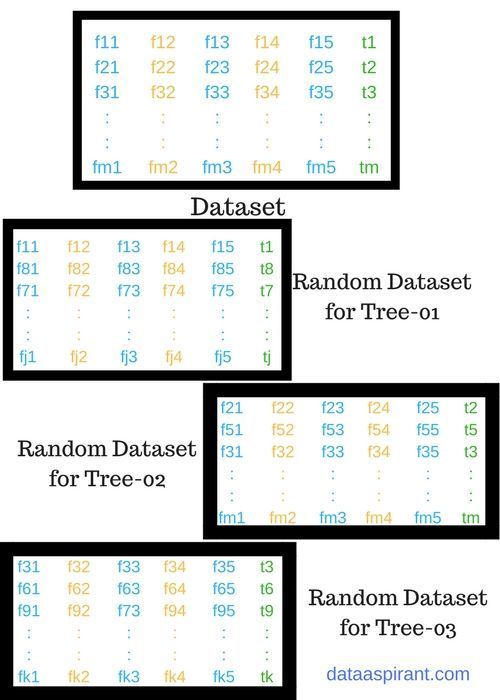
Testing of data is done based on training model which is classified using supervised learning algorithm. Evaluation of the total responses for every question and determine the polarity of feedback received in context of the given data.

## 4.1.1. Figures and table





# 4.2 Statistical techniques and visualization



How random forest algorithm works

Let’s look at the pseudo code for random forest algorithm and later we can walk through each step in the random forest algorithm.

The pseudo code for random forest algorithm can split into two stages.

* **Random forest** creation pseudo code.
* **Pseudo code to perform prediction** from the created random forest classifier.

First, let’s begin with random forest creation pseudo code

#### Random Forest pseudo code:

1. Randomly select **“k”** features from total **“m”** features.
   1. Where **k << m**
2. Among the**“k”** features, calculate the node **“d”** using the best split point.
3. Split the node into **daughter nodes** using the **best split**.
4. Repeat **1 to 3** steps until “l” number of nodes has been reached.
5. Build forest by repeating steps **1 to 4** for “n” number times to create **“n” number of trees**.

#### Random forest prediction pseudo code:

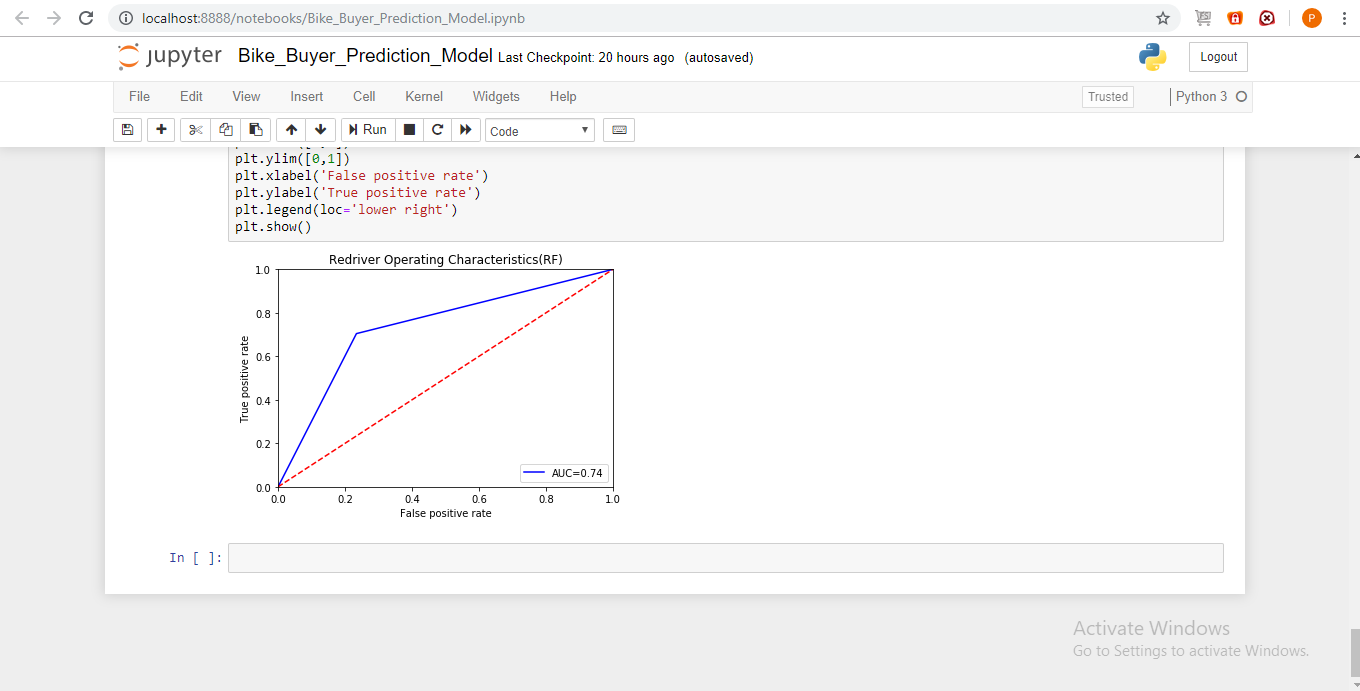
To perform prediction using the trained random forest algorithm uses the below pseudocode.

1. Takes the **test features** and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target)
2. Calculate the **votes** for each predicted target.
3. Consider the **high voted** predicted target as the **final prediction** from the random forest algorithm.

This concept of voting is known as **majority voting**.

# 4.3 Data Modeling and visualization

From sklearn.ensemble we import the RandomForestClassifier and we assign "rf" variable with RandomForestClassifier and then we give the splitted values to the model with "fit()" to predict the values we use predict(). Here we are giving inputs as gender, income, no. of children, home owner, cars owned, age and we get output as Bike buyer.



# 6. Conclusion

With the help of machine learning technology it has become easy to find out relation and patterns among various data. The work in this project mainly contains the prediction of bike buyer based on the existing customers dataset. By using Random Forest classifier algorithm we have built this model.