# Mobile Price Range Prediction(Classification)

**Ranjith K**
**Data science trainee,**
**AlmaBetter, Bangalore**

## Abstract:

As a greater number of phones are releasing in the market with lot of new features making it a brainstorm for both customer point of view and manufacturer point of view. This project helps us to infer the demand-supply information in the market and thus reducing the burden of complication for the manufacturing company to understand the price distribution for various features of their Phones.

## 1.Problem Statement:

Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

### Variables

Battery_power - Total energy a battery can store in one time measured in mAh

Blue - Has bluetooth or not

Clock_speed - speed at which microprocessor executes instructions

Dual_sim - Has dual sim support or not

Fc - Front Camera mega pixels

Four_g - Has 4G or not

Int_memory - Internal Memory in Gigabytes

M_dep - Mobile Depth in cm

Mobile_wt - Weight of mobile phone

N_cores - Number of cores of processor

Pc - Primary Camera mega pixels

Px_height - Pixel Resolution Height

Px_width - Pixel Resolution Width

Ram - Random Access Memory in Mega Bytes

Sc_h - Screen Height of mobile in cm

Sc_w - Screen Width of mobile in cm

Talk_time - longest time that a single battery charge will last when you are
Three_g - Has 3G or not

Touch_screen - Has touch screen or not

Wifi - Has wifi or not

Price_range - This is the target variable with value of 0(low cost), 1(medium cost),

2(high cost) and 3(very high cost).

## 2. Introduction

The present scenario is about how good is the customer service in any industry as the number of options at the customer's disposal are unlimited. So, it becomes extremely important to make sure that the customers would get the desired phones under their budget. It would also not be practical to keep lot of phones even when it is outdated and the demand is low. Hence, with the help of machine learning, this project aims at predicting the price range of phones so that it provides solutions for manufacturer and as well as consumer

## 3. Steps involved:

- **Exploratory Data Analysis:**

  The first step of our project is performing the EDA process on the dataset so that we can get the idea about the dataset i.e., the number of variables, the data type of the variables, visualize the dataset for better understanding and decide the suitable methods and algorithms that might produce desired Outcomes.

- **Data Preprocessing:**

  The dataset was imported and read the csv file, null values were taken care of at first, Distribution was created using histogram for numerical features, later outliers were found out with the help of box plot and treated by log transformation.
  Label encoding of categorical values was done. Correlation heat map was generated to understand the correlation among the variables and removed the features which has high correlation.

- **Building Machine Learning Model:**

After the data preprocessing is done then the data will be ready to be fit into machine learning models. We have used algorithms such as Logistic Regression,

K Nearest Neighbors, XGBoost Classifier for the prediction and used Regression Evaluation Metrics to find out the best fit models.

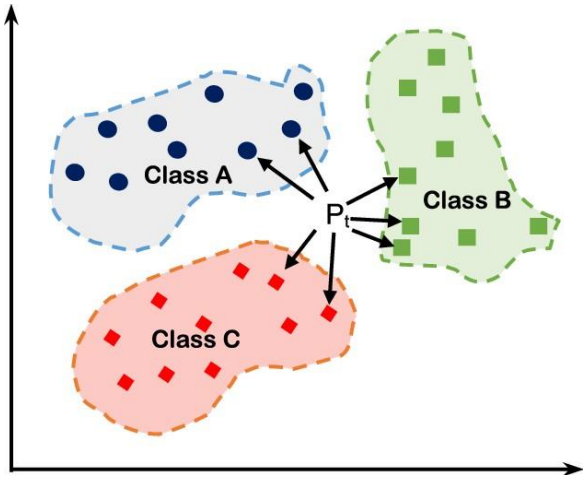
- **Summary:**

At last the summary of the project is described to have brief look over the project.

# 4.Algorithms:

## 1.K Nearest Neighbors:

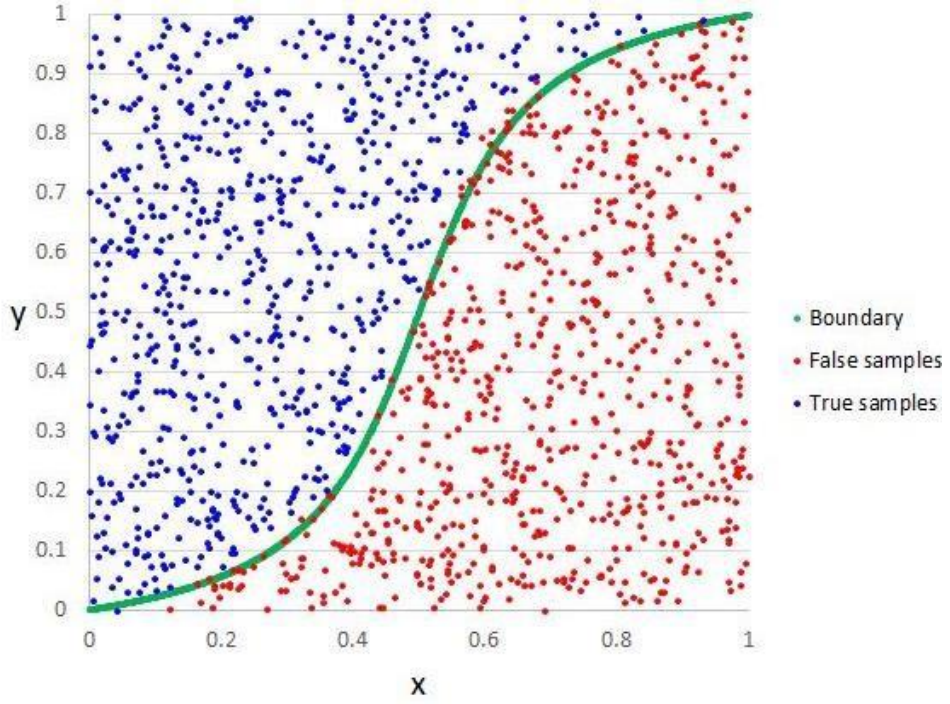
- K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.




Fig(a): Example of KNN

## 2.Polynomial Regression Model:

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas **Logistic regression is used for solving the classification problems**.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).



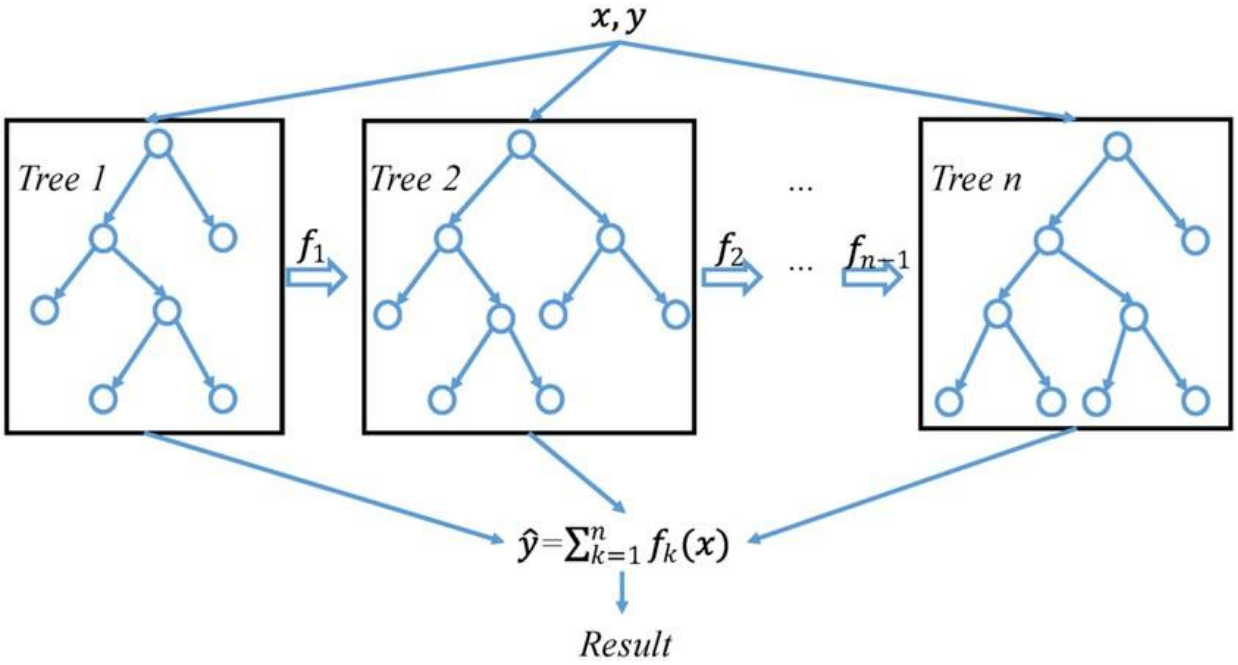

Fig(b): Example of polynomial regression.

### 3. XGBoost Classifier:

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks.

The algorithm differentiates itself in the following ways:

1. A wide range of applications: Can be used to solve regression, classification, ranking, and user-defined prediction problems.
2. Portability: Runs smoothly on Windows, Linux, and OS X.
3. Languages: Supports all major programming languages including C++, Python, R, Java, Scala, and Julia.
4. Cloud Integration: Supports AWS, Azure, and Yarn clusters and works well with Flink, Spark, and other ecosystems.



## 5. Hyper parameter tuning:

### Randomized Search CV:

RandomizedSearchCV implements a “fit” and a “score” method. It also implements “score_samples”, “predict”, “predict_proba”, “decision_function”, “transform” and “inverse_transform” if they are implemented in the estimator used.

The parameters of the estimator used to apply these methods are optimized by cross-validated search over parameter settings.

In contrast to GridSearchCV, not all parameter values are tried out, but rather a fixed number of parameter settings is sampled from the specified distributions. The number of parameter settings that are tried is given by n_iter.

If all parameters are presented as a list, sampling without replacement is performed. If at least one parameter is given as a distribution, sampling with replacement is used. It is highly recommended to use continuous distributions for continuous parameters.

## 7.Conclusion:

1. The 'price range' of the given dataset has equal distribution of the total number of phones in each of the price range with 500 nos.
2. It isobserved that 76.2 percent are 3g supported and 27.8 percent are not supported.
3. It is observed that 52.1 percent are 4g supported and 47.9 percent are not supported.
4. There is only few numbers of outlier in 'fc' column in feature engineering and we can neglect it as it has negligible amount.
5. During Multivariate analysis, in correlation heatmap, we get to see that 'ram is highly correlated with 'price range' thus inferring that 'ram' has high impact on price prediction.

6. In K nearest Neighbors classification model, we have got the knn score as 56.25%, accuracy score of 67% for training set and 65% for test set.
7. During 'elbow method' we have got the insight that the optimum value of k is 22 with least error rate.
8. In Logistic Regression Model, we have got the log score as 91%.
accuracy score of 98% for training set and 91% for test set.
9. In XGBoost model the score was 89% before hyper parameter tuning.
10. RandomizedSearchCV is used for hyperparameter tuning in XGBoost classifier and the accuracy obtained after hyper parameter tuning was 86% for training set and 80% for test set.
11. Finally, in the model explaininabilty we have used shap and we got the ins ight that 'ram', 'battery power', and phone dimensions are the features which is deciding as key factor for the price range prediction.