MOBILE PRICE RANGE PREDICTION

Pratyush Kumar Rath
Gaurav Singh
Maharana Bikesh
AlmaBetter, Bangalore

ABSTRACT:

The main motive of this research work is to predict "whether a mobile phone with given features will be economical or expensive". Various feature selection algorithms are used to identify and remove less important and redundant features and have minimal computational complexity. Various classifiers are used to achieve the highest possible accuracy. The results are compared in terms of the highest achieved accuracy and minimum selected properties. The conclusion is made based on the best feature selection algorithm and the best classifier for the given data set. This work can be used in any type of marketing and business to find the optimal product (with minimum cost and maximum features). In the future, it is proposed to expand this research and find a more sophisticated solution to the given problem and a more accurate tool for price estimation.

Keywords-, exploratory data analysis, Price prediction, feature engineering, Machine Learning, Decision Tree, Naïve Bayes, Data mining , Data Cleaning.

INTRODUCTION:

Price is the most effective marketing and business attribute. The very first question of the customer concerns the price of the goods. All customers are concerned at first and think “Whether he would be able to buy something with the given specifications or not”. So estimating the price at home is the basic purpose of the job. This document is only the first step towards the above goal.

Artificial intelligence – making a machine able to answer questions intelligently – is a very large field of engineering today. Machine learning gives us the best techniques for artificial intelligence like classification, regression, supervised and unsupervised learning and many more. There are various tools available for machine learning tasks like MATLAB, Python etc. We can use any of the classifiers like Decision tree , Naïve Bayes and many more. Different types of feature selection algorithms are available to select only the best features and minimize the dataset. This reduces the computational complexity of the problem. Since this is an optimization problem, many optimization techniques are also used to reduce the dimensionality of the dataset.

When estimating the price of a mobile phone, it is very important to consider many features. For example, a mobile phone processor. Battery timing is also very important in today's busy schedule of human being. The size and thickness of the mobile are also important factors in the decision. Internal memory, camera pixels and video quality should be taken into consideration. Internet browsing is also one of the most important limitations of this technological era of the 21st century. And so the list of many features is based on them, the mobile price is decided. So we will use many of the features above to classify whether a mobile will be low_cost, medium_cost, high_cost or very_high.

PROBLEM STATEMENT:

Maximize: The sales of Mobile

Minimize: Minimise the price of Mobile with the features.

The main goal of the project is:

In the competitive mobile phone market companies want to understand sales data of mobile phones and factors which drive the prices.The objective is to find out some relation between features of a mobile phone(eg:- RAM,Internal Memory, etc) and its selling price. In this problem, we do not have to predict the actual price but a price range indicating how high the price is.

DATA DESCRIPTION:

Here We Have A Total Of 2000 Rows And 21 Columns. And the data set includes

All Features Are Numerical Features, Among Which Features Like Battery Power, Mobile Weight, Pixel Height, Pixel Width, Ram Are Continuous Variables Others Are Discrete. Among The Discrete Variables There Are Categorical Variable Like 'Bluetooth', 'Dual Sim', '4g', '3g' ,'Wifi', 'Touch Screen'.

DATA INSIGHTS:

- There Is No Missing Values, No Null Values, No Duplicate Values.
- We Have Dependent Variable 'Price Range' Which We Will Need To Predict For Future Observations.
- It Is Basically A Classification Problem Having Classes 0, 1, 2, 3, Which Predict The Ranges Of The Mobile Price.
- First We Have To Convert Every Feature Name According To Its Corresponding Name Which Will Help Us For Better Understanding Of The Dataset.
- Here Are Our New Feature Names : 'Battery power', 'Bluetooth', 'CPU Speed', 'Dual Sim', 'Front Camera', '4G', 'ROM', 'Height', 'weight', 'cores', 'Back Camera', 'Pixel Height', 'Pixel width', 'RAM', 'Screen length', 'Screen width', 'Talk time','3G','Touch Screen', 'WiFi', 'Price Range']

Data-set description

Feature Name	Type
Battery_power -	Int64
Blue -	Int64
Clock_speed -	Float64
Dual_sim -	Int64
Fc -	Int64
Four_g -	Int64
Int_memory -	Int64
M_dep -	Int64
Mobile_wt -	Int64
N_cores -	Int64
Pc -	Float64
Px_height -	Float64
Px_width -	Int64
Ram -	Int64
Sc_h -	Int64
Sc_w -	Int64
Talk_time -	Int64
Three_g -	Int64
Touch_screen -	Int64
Wifi -	Int64
Price_range -	Int64

FEATURE BREAKDOWN:

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

Touch_screen - Has touch screen or not

Wifi - Has wifi or not

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

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).

EXPLORATORY DATA ANALYSIS:

To explain EDA simply, it means trying to understand the given data much better in order to make some sense out of it. Using univariate frequency analysis, we described the key characteristics of each element including minimum and maximum value, mean, standard deviation, and more. It was also used to create a distribution of values and identify missing values and outliers.
EDA is the process of examining an available data set to discover patterns,

detect anomalies, test hypotheses, and validate assumptions using statistical measurements. In this chapter, we'll discuss the steps involved in performing cutting-edge exploratory data analysis
In statistics, A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing tasked in Python uses data visualization to draw meaningful patterns and insights

STEPS EVOLVED IN DATA CYCLE

1. **DATA ANALYSIS:**
2. **DATA SOURCING**
3. **DATA PREPROCESSING:**
4. **DATA CLEANING**
5. **DATA TRANSFORMATION:**
6. **DATA DEDUPLICATION:**
7. **MISSING VALUES:**
8. **DROPPING MISSING VALUES:**
9. **HANDLING OUTLIERS:**

ALGORITHMS:

1. LINEAR REGRESSION:

Linear regression is a supervised machine learning model mostly used in forecasting. Supervised machine learning models are those where we use training data to build a model and then test the accuracy of the model using a loss function.

Linear regression is one of the most well-known time series forecasting techniques used for predictive modeling. As the name suggests, it assumes a linear relationship between a set of independent variables and a dependent variable (variable of interest).

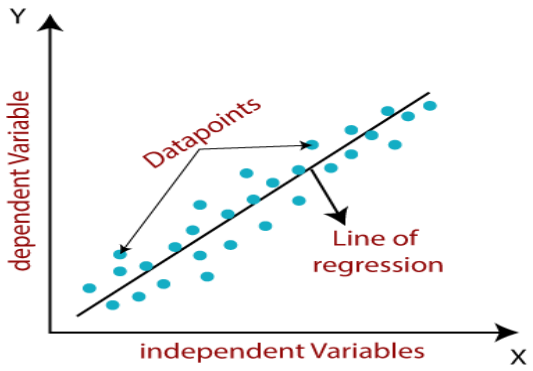
EQUATION OF BEST FIT LINE.

$$y = \beta 0 + \beta 1x$$

to our data. Here x is called the independent variable or predictor variable and y is called the dependent variable or response variable. Before we talk about how to perform the fit, let's take a closer look at the important quantities from the fit:

• β1 is the slope of the line: this is one of the most important quantities in any linear regression analysis

• β0 is the intersection of the line.

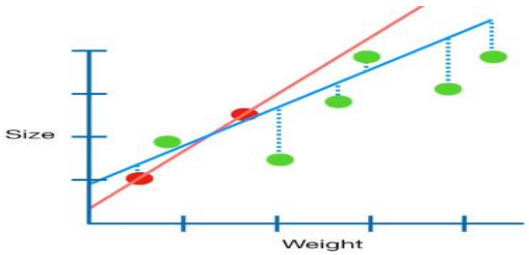


2. RIDGE REGRESSION:

Ridge regression is a model tuning method used to analyze any data that suffers from multicollinearity. This method performs L2 regularization. When the problem of multicollinearity occurs, the least squares are unbiased and the variances are large, resulting in the predicted values being far from the true values.

we came to the conclusion that we would like to reduce the complexity of the model, i.e. the number of predictors.

We could use forward or backward selection to do this, but that way we wouldn't be able to say anything about the effect of the removed variables on the response. Removing predictors from the

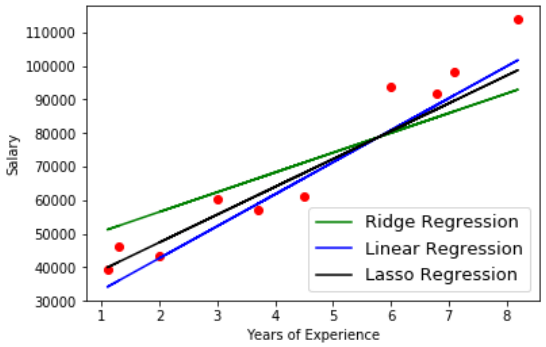


model can be thought of as setting their coefficients to zero. Instead of forcing them to be exactly zero, let's penalize them if they are too far from zero, thus forcing them to be small all the time. This way we reduce the complexity of the model while keeping all the variables in the model. This is essentially what Ridge Regression does.

3. LASSO REGRESSION:

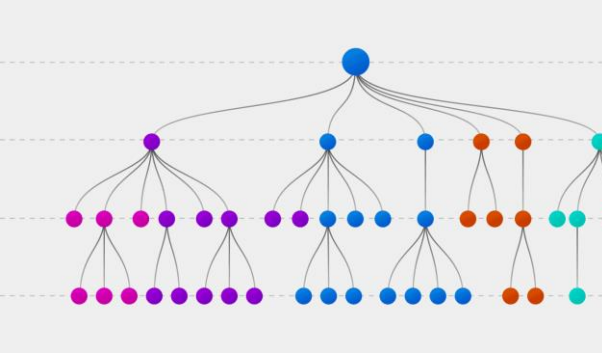
Lasso, or the least absolute shrinkage and selection operator, is conceptually quite similar to ridge regression. It also adds a penalty for nonzero coefficients, but unlike ridge regression, which penalizes the sum of the squares of the coefficients (the so-called L2 penalty), lasso penalizes the sum of their absolute values (the L1 penalty). As a result, for high values of λ, many coefficients are exactly zero under the lasso, which is never the case for ridge regression.

The only difference in the comb and lasso loss functions is in the penalty conditions. Under the lasso, loss is defined as:



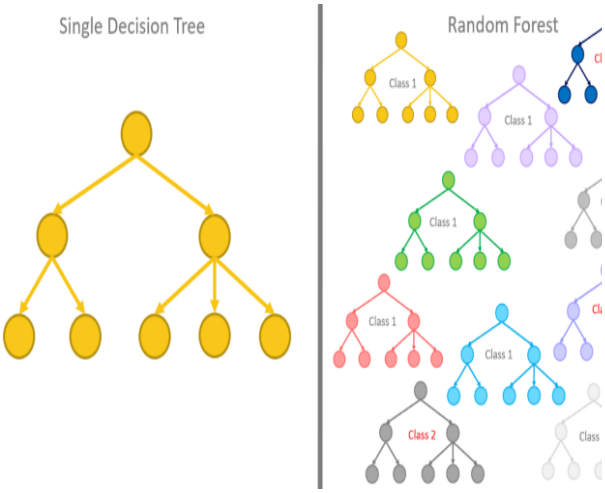
4.DECISION TREE:

A decision tree is the most powerful and popular tool for classification and prediction. A decision tree is a tree structure similar to a flowchart, where each internal node denotes a test on an attribute, each branch represents the result of the test, and each leaf node (terminal node) has a class label. A tree can be "learned" by dividing the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. Decision trees classify instances by ordering them in a tree from the root to some leaf node that provides the instance's classification. An instance is classified by starting at the root node of the tree, testing the attribute specified by that node, and then moving down the branch of the tree corresponding to the value of the attribute, as shown in the figure above. This process is then repeated for the subtree rooted at the new node.



5. RANDOM FOREST:

Random Forest is a wrapper type of decision tree algorithm that creates a series of decision trees from a randomly selected subset of the training set, collects labels from those subsets, and then averages the final prediction depending on how many times the label has been used. predicted of all.



6. GRADIENT BOOSTING:

The term gradient gain consists of two sub-terms, gradient and gain. We already know that gradient boosting is a boosting technique. Let's see how the term "gradient" relates here.

Gradient boosting redefines boosting as a numerical optimization problem where the objective is to minimize the loss function of the model by adding weak pupils using gradient descent. Gradient descent is a first-order iterative optimization algorithm for finding the local minimum of a differentiable function. Since gradient boosting is based on the minimization of a loss function, different types of loss functions can be used, resulting in a flexible technique that can be applied to regression, multi-class classification, etc.

CONCLUSIONS:

- From The EDA We Get That The Feature 'RAM' Is The Most Relative Feature With Our Dependable Feature 'Price Range'. More RAM Capacity Will Increase The Price Of The Mobile.
- By Implementing All The Models We Have Found That Gradient Boosting Has The Best Score Of 91.2% Followed By Logistic Regression of 91.1% And Random Forest Of 89%.
- By Appling Hyper Parameter Tuning We Get The Score Of Gradient Boosting Of 92%.
- The Battery Power And Pixel Height Also Positively Correlated With Price Range That Means It Will Also Create Some Impact To The Calculation.
- So We Can Conclude That The Best For This Dataset Is Gradient Boosting With Hyper Parameter Tuning, Which Can Be Used For Future Prediction Of The DataSet.

- https://book.akij.net/eBooks/2018/May/5aef50939a868/Data_Science_for_Bus.pdf
- https://bunker2.zlibcdn.com/dtoken/01c5fc197a94283bfb0c0943bd5b2d0c
- W3SCHOOLS
- GEEKS FOR GEEKS
- ALMA BETTER
- PANDAS DOCUMENTATION

REFERENCES:

- Data science for business : what you think about data mining