Mahavir Education Trust's

SHAH & ANCHOR KUTCHHI ENGINEERING COLLEGE

Chembur, Mumbai - 400 088

UG Program in Computer Engineering

Mini Project Report

Academic Year: 2022-2023

Semester: 7 **Subject: Machine Learning Lab**

Aim	To predict phone price range with best accuracy by analyzing features of the phone.
Project Title	Phone Price Prediction

Group Members			
Class	Batch	Roll No.	Name
BE3	C	42	Darshan Shyamkant Patil
BE3	C	47	Varad Nitin Potdar
BE3	C	50	Chintan Jayesh Rajgor

Table of Contents:

Introduction

Price is the most powerful marketing and commercial feature. The first enquiry from a customer is regarding the pricing of the things. All of the customers are first concerned and wonder, "Will he be able to acquire something with the supplied criteria or not?" So the primary goal of the task is to estimate prices at home. This study is merely the first step in reaching the aforementioned goal.Nowadays, mobile phones are among the most popular selling and purchasing devices. Every day, new mobile phones with new versions and enhanced functions are released. Each day, hundreds of thousands of mobile phones are sold and purchased. So, in this situation, the mobile price class prediction is a case study for the specified type of issue, namely, identifying the best product.

Literature review

The use of historical data to forecast the pricing of available and new launch products is an intriguing study background for machine-learning researchers. During investigation, it was discovered that most standard algorithms, such as Decision Tree and Nave Bayes, are incapable of processing, categorizing, and forecasting numeric data. Mariana Listiani, a researcher, uses the notion of supporting vector machine (SVM) in her work predicting the pricing of used vehicles. When there is a big data set, it has been discovered that the SVM approach is significantly better and more dependable for price prediction than other methods such as multiple linear regressions. The study also demonstrated that SVM is useful for handling high-dimensional data and minimizing both under-fitting and overfitting. Genetic Algorithm used to identify essential features for SVM Listiani.

Objective

The main objective of this project is to study and understand dataset with the help of the machine learning algorithms and draw out the predictions of the phone price ranges. This will be beneficial to find if the phone is economically expensive or reasonably priced. This can help a lot of industries directly and indirectly linked by mobile phones and hence a huge profit can be taken by these industries. This can give insight to which component or which feature is more efficient

Comparisons of methods

1.Linear Regression:

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x)

$$h_\theta = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots$$

Loss function :In LR, we use mean squared error as the metric of loss. The deviation of expected and actual outputs will be squared and sum up. Derivative of this loss will be used by gradient descend algorithm.

2.Logistic Regression:

logistic regression (or logit regression) is estimating the parameters of a logistic model (the coefficients in the linear combination)

Loss function **:**

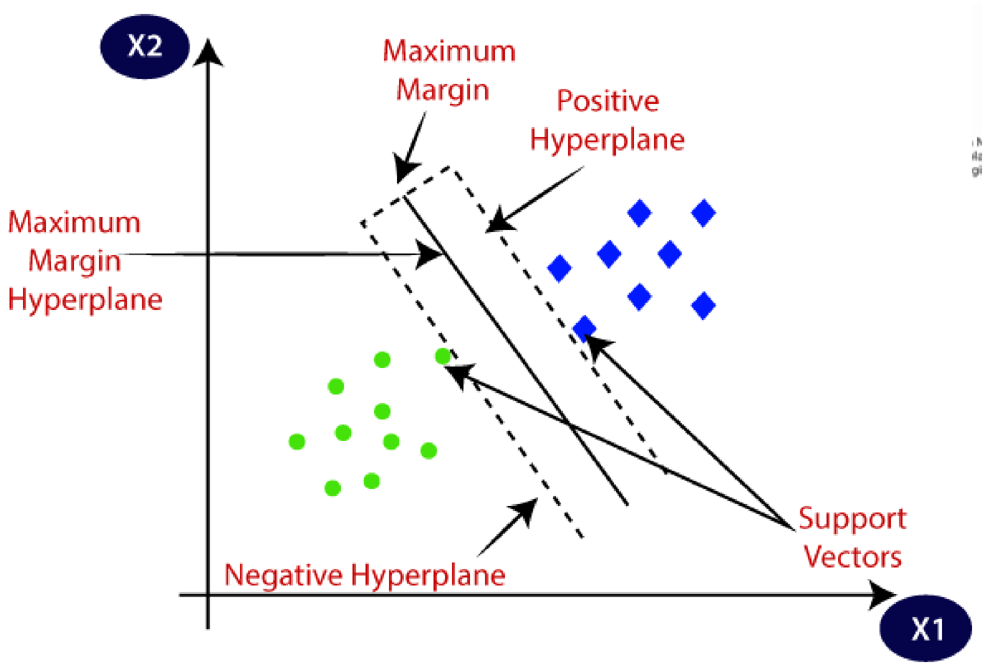
$$J(\theta) = \frac{1}{m}\sum cost(y', y)$$

$$cost(y', y) = -log(1 - y') \;\; if \;\; y = \;0$$

$$cost(y', y) = -log(y') \;\; if \;\; y = 1$$

3.SVM(Support Vector Machine):

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection using support vectors and creating hyperplanes with distance between them(Maximum Margin).



4.DT(Decision Tree):

A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

for CART(classification and regression trees), we use gini index as the classification metric. It is a metric to calculate how well the datapoints are mixed together.

$$giniindex = 1 - \sum P_t^2$$

Logistic regression vs SVM :

- SVM can handle non-linear solutions whereas logistic regression can only handle linear solutions.
- Linear SVM handles outliers better, as it derives maximum margin solution.
- Hinge loss in SVM outperforms log loss in LR.

Logistic Regression vs Decision Tree :

- Decision tree handles colinearity better than LR.
- Decision trees cannot derive the significance of features, but LR can.
- Decision trees are better for categorical values than LR.

LR vs Decision Tree :

- Decision trees supports non linearity, where LR supports only linear solutions.
- When there are large number of features with less data-sets(with low noise), linear regressions may outperform Decision trees/random forests. In general cases, Decision trees will be having better average accuracy.
- For categorical independent variables, decision trees are better than linear regression.
- Decision trees handles colinearity better than LR.

LR vs SVM :

- SVM supports both linear and non-linear solutions using kernel trick.
- SVM handles outliers better than LR.
- Both perform well when the training data is less, and there are large number of features.

Decision tree vs Random Forest :

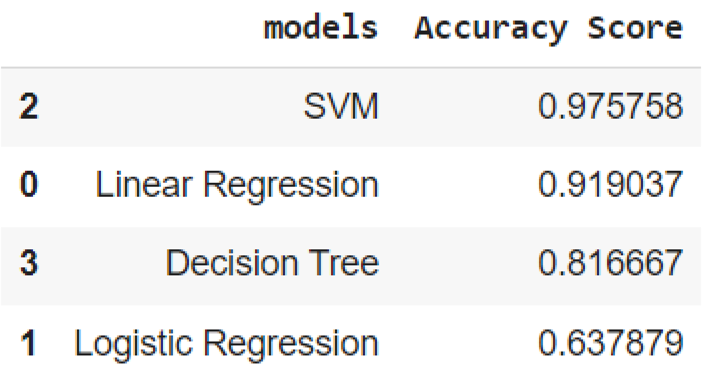
- Random Forest is a collection of decision trees and average/majority vote of the forest is selected as the predicted output.
- Random Forest model will be less prone to overfitting than Decision tree, and gives a more generalized solution.
- Random Forest is more robust and accurate than decision trees.

Decision tree vs SVM :

- SVM uses kernel trick to solve non-linear problems whereas decision trees derive hyper-rectangles in input space to solve the problem.
- Decision trees are better for categorical data and it deals colinearity better than SVM.

Results

After training all the models with the same dataset the testing accuracy of the predicted output is as follows:



	models	Accuracy Score
2	SVM	0.975758
0	Linear Regression	0.919037
3	Decision Tree	0.816667
1	Logistic Regression	0.637879

Therefore from the results we can analyze which model is more suitable for dataset of phone prices and helps in providing better predictions.

Conclusion

Cost estimating is an important part of marketing and business. Finding the best product is the finest marketing approach (with minimum cost and maximum specifications). Thus, products may be measured in terms of their needs, pricing, Production Company, and so on. A good product may be recommended to a customer by establishing the economic range, which is best accomplished through data mining and research. The price range of a mobile phone was successfully predicted with high accuracy in our use case by training the model on a dataset of two thousand cases with varied variables. The best accuracy for the dataset to provide accurate predictions is SVM(Support Vector Machine). Hence we can use the SVM model to predict the unknown data and get the predicted result of the phone price range.

References

- https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm
- https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222
- https://towardsdatascience.com/a-complete-view-of-decision-trees-and-svm-in-machine-learning-f9f3d19a337b
- https://rjunaidraza.medium.com/comparison-of-classification-algorithms-lr-dt-rf-svm-knn-6631493e300f
- https://en.wikipedia.org/wiki/Support_vector_machine
- https://iwaponline.com/hr/article/48/5/1214/1588/A-comparative-study-of-multiple-linear-regression
- https://www.geeksforgeeks.org/differentiate-between-support-vector-machine-and-logistic-regression/