Exploratory Data Analysis and Customer Segmentation for Smartphones

Ritika Singh[1], Sunidhi[2], Suryansh[3]

[1]*Assistant Professor, Department of CSE, SRM Institute of Science and Technology, Ghaziabad*

[2]*Department of CSE, SRM Institute of Science and Technology, Modinagar, Ghaziabad*

[3]*Department of CSE, SRM Institute of Science and Technology, Modinagar, Ghaziabad*

ABSTRACT

In today's modern and technological era, the smartphones are emerging with new innovative features every time in order to satisfy consumer's professional and personal needs. The research is directed towards exploring the features which is more influential than others for predicting the overall cost of the device in order to give idea whether the phone would be economical or expensive. The work uses the knowledge of python programming language, knowledge of various classification algorithms to classify and predict price. Results are compared in terms of highest accuracy achieved between the model built with a lazy algorithm and the other having eager learning algorithm. The overall objective of the project is to build a model which segments customers on the basis of price-range they fall into according to their required specifications in a phone. This type of research may be very helpful in marketing and business sectors or anyone who intends to find an optimal product.

General Terms

Data Science & Analytics, Machine Learning

Keywords- Data Preparation, Descriptive & Inferential Statistics, Feature Selection, KNN, Random Forest

1. INTRODUCTION

Customer Segmentation is all about categorizing the customers on the basis of some similar properties and grouping them collectively. Segmenting users is one of the key aspects in the mobile industry and thus targeting the right consumers. We intended our work towards segmenting users on the basis of price.

Cost is the best property of showcasing and business. The absolute first inquiry of costumer is about the cost of things. Every customer is concerned that he would have the option to buy something with given specifications or not.

Prediction of the prices can be achieved by building a ML model using classification and supervised learning process under this field. We can use any of algorithms like, Decision Tree, KNN, Random forest etc. Different type of feature selection algorithms are available like select k best, feature importance to select only best features and minimize dataset. This reduces computational complexity of the problem. But before approaching towards Machine Learning problems, it is better to study and understand the gathered Dataset. The analysis of Data helps to inference insights and also makes the data more feasible and easy to deal with.

Smartphone these days is quite possibly the most trending and demanding gadget. Consistently new mobiles with new form and more features are created and brought in the market.

Countless numbers of phones are marketed on regular basis. Hence this project is concentrated on finding the efficient device according to the people's need and affordability. A similar work should be possible to appraise cost of all items like which are meant to be grouped into price categories, like food, vehicles etc.

Various specifications of a phone collectively contribute in the price estimation. The data which is being researched on consists of many features which includes Cust ID, Battery power, Bluetooth availability, Clock speed, Dual sim or single sim, Front camera quality, having four G or not, Internal memory, Depth in cm, Device color, Weight, Number of cores of processor, Pixel height and width, Power of RAM etc. Main Goal is to thus classify the phone's price category (target variable) with the help of these features.

2. LITERATURE SURVEY

[1] Shubhiksha S., Swathi Thota, J. Sangeetha Predicted the Phone prices using phone's different specifications as dependent factors of cost. In their research they used Random Forest Classifiers, SVM and Logistic regression where they found SVM as most accurate 81 percent. They solved the problem by taking historical data pertaining to the key features of smartphones along with cost to build predictive model. [2] Sameerchand-Pudaruth predicted the prices of second hand cars. He implemented various classifiers like Multiple linear regression, (KNN), Decision Tree, and Naive Bayes to predict the prices. During research it was found that Decision Tree and Naive Bayes proved to be inefficient because of the lower Number of instances for his research.[3] Pritish Arora, S. Srivastava, B.Garg carried their experiment of mobile price prediction using WEKA. On the other hand, [4] Dineshkumar E predicted house prices using machine learning. He used Linear Regression to achieve his goal.
If we throw light on 'Segmentation' then, [5] DiFadly Hamka's paperwork explores the use of market segmentation on the perspective of actors in mobile ecosystem which are network operator, handset manufacturer and application provider. He also explored the interaction that may exist between each actor by analysing the relation between the resulted segments of each perspective. [6] Shonda Kuiper discussed the development of multivariate regression model to predict the retail price of 2005 General Motors Cars.

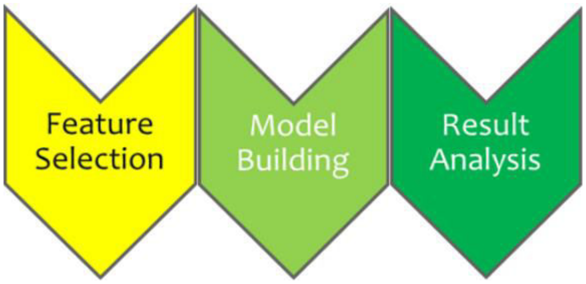
There is a lot of scope to predict and compare price segmentation models with more available algorithms and by using sufficient size of validated data

.

3. METHODOLOGY

a) Exploratory Data Analysis:

1- Data preparation
2- Descriptive statistics
3- Inferential statistics



b) Model Building

1- Feature Importance, Select K Best
2- KNN and Random Forest
3- Measuring accuracy by evaluating confusion matrix.

3.1DATA ANALYSIS

Data Preparation

The validated dataset consisting of features of mobiles are collected from https://www.kaggle.com. For Data Editing, the dataset is carefully checked for the presence of any null value. It is also checked for redundant rows and removal of unnecessary columns (‘ID’, ‘color’) is done carefully. Rest, the dataset was already very precise and there was no need for any kind of data encoding.

Descriptive Statistics

Univariate analysis helped in understanding the dispersion and shape of the collected dataset’s distribution. It summarizes the data clearly. Here, the used dataset has 1973 data entries and 21 columns including the target variable- ‘price range’ where 0(lowest range), 1(medium cost), 2(high cost), 3(expensive range). Description also made aware of the dimension, shape, size and memory usage of the dataset.

Inferential Statistics

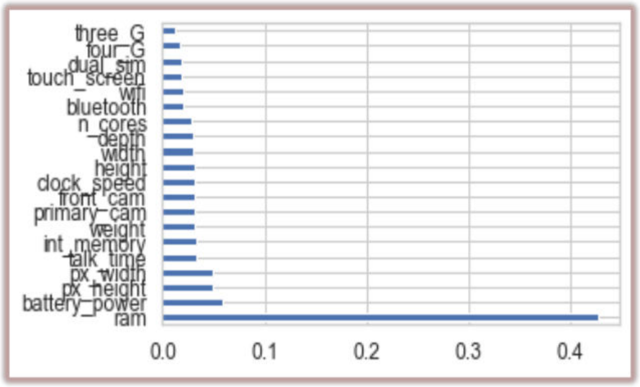
Helps in making generalizations, identifying patterns and correlations. This is used to draw the insights and not for the actual scaling. Histograms, Joint plots, Point plots, Line plots, box plots, Overlaying plots, Bar plots, Pie plots are used to visualize relations and value distributions. RAM, Battery power, Talk time, Pixel height, Clock speed and Front camera came out as the most contributing features for price.

3.2 MODEL BUILDING

Feature Selection

Feature selection algorithms is all about confirming our drawn insights and telling the real important features for the model building. Feature selection by embedded methods considers set of all possible subsets of categories and select the best subset responsible for predictions and learning of algorithms. We have used two selection methods. One is Select k Best method which uses chi2 statistical test. The other is Feature importance method which calculates the power of each features of the dataset.

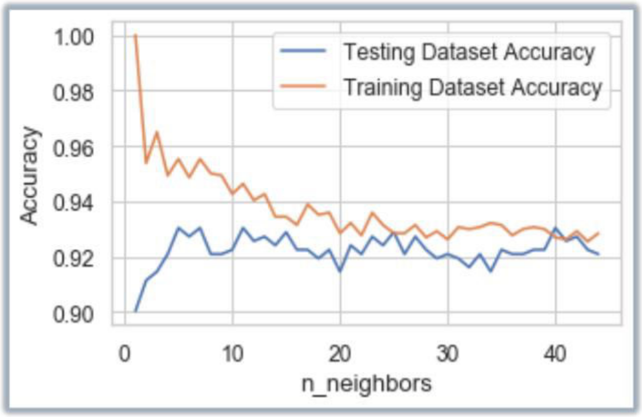
After applying feature selection algorithms, it is now confirmed that the top features for price prediction are :- ram , battery_power , px_width , px_height , weight , int_memory. This is shown below:



Training Models

We have used one lazy learner classifier and the eager learner classifier to build two predictive models to predict the cost ranges of smartphones.

K Nearest is a simple but very powerful algorithm that stores all the available cases and classifies the new data or case based on a similarity measure. KNN is one of the lazy learning algo and one of the biggest use of KNN is a recommender system. One of the most important step in building the model using KNN is to find the right K value for the dataset. Generally the right k value lies near to the under root of n where n is the number of data entries. Our procedure for finding the right K is illustrated below which shows that the best suited value for K is 40, giving the highest accuracy for the testing dataset at 93%



Another algorithm used to build model in our work is Random Forest Classifier to achieve multi class classification using Scikit-Learn Library. This algorithm is one of the eager learners. This classifier does not rely on singular decision, it assembles randomized decisions based on several decisions and makes the final decision based on majority. It is like the collection of multiple decision trees. The steps we followed for building the model are: Importing the required modules, Splitting the data into test/train sets, fitting the training dataset into the imported model , predicting the test set results, comparing the values of the test set before and after prediction and calculating the accuracy of the built model.

ActualClassification	***PredictedClassification***
0	**0**
1 1	
3	**3**
3 3	
1	**2**
……	
2	**2**
0 0	
1	**1**
2 2	
0	**0**

632 rows × 2 columns

4.RESULTS

4.1 Random Forest Model

```
Precision     recall  f1-score  support

                 0   0.91       0.94
0.92             169
                 1   0.80       0.83
0.81             167
                 2   0.71       0.67
0.69             150
                 3   0.84       0.82
0.83             146

Accuracy
0.82             632
macro avg        0.81       0.81
0.81             632
weighted avg     0.82       0.82
0.82             632

<function     confusion_matrix
at 0x000000000B30CEE8>
---------
the   score   of   the   Random   Forest
model is: 0.8180379746835443
```

4.2 KNN Classification Model

```
Precision     recall  f1-score  support

                 0     0.96      0.99
0.97             169
                 1     0.90      0.90
0.90             167
                 2     0.89      0.88
0.88             150
                 3     0.97      0.95
0.96             146

Accuracy
0.93             632
macro avg      0.93 0.93
0.93             632
weighted avg         0.93      0.93
0.93             632

<function     confusion_matrix
at 0x000000000B30CEE8>
---------
the score of the knn model is:
0.930379746835443
```

Best accuracy model for our research came out to be the KNN classification model which showed a total accuracy of 93% .

5. CONCLUSION

After analyzing the ambiguous and raw collected data we are able to clean, sanitize and reduce it to more sensible form. We found various correlations and patterns between various dependent and independent variables. Our most important insight derived was to identify dominant features for the prediction of price and dividing customer's choices into economical categories. Since the classification was supervised learning, the algorithm model is trained by giving well targeted data.

Main goal was to identify the correct category or class to which a new data will fall under. Goal is achieved by building a price prediction model using KNN algorithm and Random Forest and evaluation report is being produced which showed overall 93% and 81% accuracy respectively.

6. FUTURE POSSIBILITIES

This sort of prediction will help organizations estimate cost of mobiles to give extreme competition to other companies. Additionally it will be valuable for Consumers to satisfy that they are buying a handset with best possible value of it.

More refined AI methods can be utilized to expanded the exactness and predict the precise cost of the items. Mobile application can be built up that will foresee the market cost of any new dispatched item.

To accomplish greatest accuracy and predict more precise, an ever increasing number of occasions ought to be added to the informational collection. Furthermore, selecting more appropriate features can also increase the accuracy. So informational collection ought to be enormous and more proper highlights ought to be chosen to accomplish higher precision**.**

7. REFERENCES

[1] S., Subhiksha and Thota, Swathi and Sangeetha. "Prediction of Phone Prices Using Machine Learning Techniques" January 2020 DOI:10.1007/978-981-15-1097-7_65 In book: Data Engineering and Communication Technology (pp.781-789)

[2] Sameer Chand Pudaruth. "Predicting the Price of Used Cars using Machine Learning Techniques", International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, Number 7 (2014), pp. 753764

[3] Pritish Arora , Sudhanshu Srivastava , Bindu Garg "Mobile Price Prediction using WEKA", International Journal of Science & Engineering Development Research (www.ijsdr.org), ISSN:2455-2631, Vol.5, Issue 4, page no.330 - 333, April-2020

[4] Dineshkumar E, "House price prediction machine learning project using python."

[5] Fadly hamka, Harry Bouwman, Markde Reuver, Maarten Kroesen ." Mobile customer segmentation based on smartphone measurement" Telematics and Informatics Volume 31, Issue 2, May 2014, Pages 220-227

[6] Shonda Kuiper. " Introduction to Multiple Regression: How Much Is Your Car Worth? "November 2008 Journal of Statistics Education 16(3) DOI:10.1080/10691898.2008.11889579

[7] Wattana Punlumjeak, Nachirat Rachburee**.** "A comparative study of feature selection techniques forclassifystudentperformance**"** DOI:10.1109/ICITEED.2015.7408984 Corpus ID: 15974391 Published 2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE)

[8] Kanwal Noor, Sadaqat Jan. "Vehicle Price Prediction System using Machine Learning Techniques" International Journal of Computer Applications (0975 – 8887) Volume 167 – No.9, June 2017

[9] How to _nd the optimal value of K in KNN?,https://towardsdatascience.com/how-to-_nd-the-optimal-value-of-k-in-knn 35d936e554eb

[10] Feature Selection Techniques in Machine Learning with Python, https://towardsdatascience.com/feature-selection-techniques-in- machine-learning-with-python-f24e7da3f36e

[11] Data analysis in research: Why data, types of data, data analysis in qualitative and quantitative research, https://www.questionpro.com/blog/data-analysis-in-research/