MOBILE PRICE CLASSIFICATION

DATA REPORT

Group Name: Wazito

Group Members

1. Stephen Njuguna - Scrum Master
2. Kevin Kilonzo
3. Ayub Bett
4. Brenda Chepkoech

1. BUSINESS UNDERSTANDING

Business Overview

The increasing use of smartphones is evident in the world today. For mobile dealers, an effective pricing strategy is essential for continued sales success. It is important to have a carefully planned pricing strategy.

Smartphone prices can be affected by a wide range of factors, such as brand and the length of time a particular device has been on the market. The ability to use a device internationally can also affect its price. If a device has a high-resolution camera and a lot of storage capacity, it will likely cost more than a unit without less notable capabilities. Those individuals who are willing to enter a contract, however, may be able to get a smartphone loaded with desirable features at a bargain price.

Business Objectives

Based on the features of the phone, we build a machine learning model that predicts its price.

Business Success Criteria

Finding the factors that affect client repayment ability and having a model that accurately predicts

Assessing The Situation

1. **Resources inventory**
   a. The dataset
      The Dataset [Link]
   b. Software

Github, R, Jira

2. Assumptions

The data is correct and up to date

3. Constraints

There were no constraints. Our data does not contain missing values and duplicates but it had outliers that we didn't drop because they were few and useful in our analysis.

Data Mining Goals

1. How battery power affects the price range
2. How to clock speed affects the price range
3. How number of cores affects the price range
4. How Bluetooth affects the price range
5. How dual sim affects the price range
6. How internal memory size affects the price range

Data Mining Success Criteria

We did an explanatory analysis in order to understand how the various factors affected the range of price for mobile phones

2. Data Understanding

Data Understanding Overview

For this project, we are using the dataset from the Kaggle website. These datasets are;

- Train Dataset - This dataset contains the different factors that are considered when determining the different mobile price ranges.

Train Dataset - This dataset contains factors that were recorded by the phone reseller.
It has 2000 rows and 21 columns.

Data Description

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 megapixels
Four_g - Has 4G or not
Int_memory - Internal Memory in Gigabytes
M_dep - Mobile Depth in cm(thickness)
Mobile_wt - Weight of the mobile phone
N_cores - Number of cores of a processor

Pc - Primary Camera megapixels
Px_height - Pixel Resolution Height
Px_width - Pixel Resolution Width
Ram - Random Access Memory in Megabytes
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

3. Data Preparation

These are the steps followed in preparing our data:

1. **Loading data**

We started by loading the train dataset to our R environment.

2. **Cleaning the data**

   We first checked for null values in our dataset and we found out that there were no null values. We also checked for duplicates and we found out that there were no duplicates in our dataset. There were outliers in the front camera megapixels, phone height columns but we did not drop them because they are useful in our analysis.

4. Analysis

1. **How four_g affects the price range**

- Low cost mobile with four_g are bought more than with without four_g
- Medium cost mobile with four_g are bought more than with without four_g
- High cost mobile without for_g are bought more than those with four_g.
- Very high cost mobile with four_g are bought more than those without.

2. **How does three_ g affects the price range**

- Mobiles that are three_g are bought less than though that are not three_g across all price ranges.

3. **How touch screen affects the price range**

- Low cost mobiles that are touch screened are bought more than those that are not touch screened.
- Medium cost mobiles that are touch screened are bought more than those that are not touch screened.

- High cost mobiles that are not touch screened are bought more than those that are touch screened.
- Very high cost mobiles that are touch screened are bought more than those that are not touch screened.

4. **How dual sim affects the price range**
- Low cost mobile phones are bought equally without a dual sim.
- Medium cost mobiles with dual sim are bought more than without.
- High cost mobile phones are bought equally without a dual sim.
- Very high cost mobile with dual sim is bought more than those without.

5. **How wifi affects the price range**
- Low-cost mobile with wifi has a small difference on how they are bought.
- Medium cost mobiles with wifi have a small difference on how they are bought.
- High-cost mobile phones with wifi have a small difference in how they are bought.
- Very high-cost mobile phones with wifi are bought more than those without.

6. **How battery power affects the price range**
- Low-cost mobile with battery power has a small difference on how they are bought.
- Medium cost mobiles with battery power has a small difference on how they are bought.
- High-cost mobile phones with battery power have a small difference in how they are bought.
- Very high-cost mobile phones with battery power are bought more than those without.

7. **How to clock speed affects the price range**
- There is no big difference in the clock speed in the different mobile price range

5. Data Modeling

a) Supervised Learning

Feature engineering

We started by performing feature selection.
Here we used the Boruta Model to determine the best features that we use in determining the mobile price range.

We decided to use LDA to confirm if our important features obtained from Boruta were correct by checking the coefficients of the features.
We started by label encoding the categorical columns in our dataset.

We then performed dimensionality reduction using LDA but since LDA assumes a normal distribution we first nomalize the data. We applied the Min-Max normalization function to numerical columns then added the target column to our normalized dataset.

We then performed dimensionality reduction using LDA and from the LDA results we note that LD1 takes 99.70% of the data and the most important features in our Boruta model had the highest coefficients so we are going to use them for modelling.

From the Boruta Model, the best features for determining the price range are

1. **Battery_power-** Total energy a battery can store in one time measured in mAh
2. **RAM**- Random Acess Memory in gigabytes
3. **Px_height -** Pixel Resolution Height
4. **Px_width** - Pixel Resolution Width
5. **Sc_h** - Screen Height of mobile in cm
6. **Sc_w** - Screen Width of mobile in cm

Building the supervised learning models

Below are the models that we used to make prediction.

1. Logistic Regression
2. Support Vector Machine
3. KNN(K Nearest Neighbours)

1. Logistic Regression.

From the Logistic Regression Model findings, the accuracy was 93%

2. Support Vector Machines

From the Support Vector Machines model findings, the accuracy was 96%

3. K Nearest Neighbours

From the K Nearest Neighbours model, the accuracy was 91%

b) Unsupervised Learning

Feature Engineering

We started by removing the label column so that we can remain with the features columns only in our dataset.

We then scaled the numerical variables so that our models can perform better, then marged the scaled data with the categorical dataset.

After scaling we created a t-sne model which is a model that is useful in dimensionality reduction. We used this dimensionality reduction method because our dataset is not linearly distributed and we have both categorical and continuous data in our dataset.

Building the unsupervised learning models

We used K-Means clustering to build our unsupervised model.

We used the K-Means clustering model because our dataset contains both categorical and numerical features.

From the model, we determined that it was possible to create clusters of the dataset which confirms the unsupervised learning model can be used to accurately predict the mobile price classes.

We did our analysis and modeling using python and the link to our full analysis can be found in this [github link]

6. Recommendations and conclusion

Conclusion

- High-cost mobile without for_g are bought more than those with four_g.
- Very high-cost mobile phones with four_g are bought more than those without.
- High-cost mobiles that are not touch screen are bought more than those that are touch screen
- Very high-cost mobiles that are touch screened are bought more than those that are not touch screened.
- High-cost mobile phones are bought equally without a dual sim.
- Very high-cost mobile phones with dual sims are bought more than those without.
- High-cost mobile phones with wifi have a small difference in how they are bought.
- Very high-cost mobile phones with wifi are bought more than those without.
- Very high-cost mobile phones with battery power are bought more than those without.
- There is no big difference in the clock speed in the different mobile price range
- Very high-cost mobile phones with battery power are bought more than those without.

The most important features in determining the price range were;

- **Battery_power-** Total energy a battery can store in one time measured in mAh
- **RAM**- Random Acess Memory in gigabites
- **Px_height -** Pixel Resolution Height
- **Px_width** - Pixel Resolution Width
- **Sc_h** - Screen Height of mobile in cm
- **Sc_w** - Screen Width of mobile in cm

From the supervised model, the best model for predicting the mobile price class range is the Support Vector Machines model with an accuracy of 96% which was better than the other models used to make predictions.

Recommendations

In order to make more profits;

- The phone seller should stock more phones of very high cost with four_g because they are bought more than the other phones.
- The phone seller should stock more phones that are of very high cost and are touch screened are because they bought more than those that are not touch screened.
- The phone seller should stock more phones that are of very high cost and with wifi are because they are bought more than those without.
- The phone seller should stock more phones that are of very high cost and with dual sim is because they bought more than those without.
- The phone seller should stock more phones that are of low-cost mobiles that are touch screen because they are bought more than those that are not touch screened.
- The phone seller should stock more phones that are of low-cost mobile phones with dual sim because they are bought equally without a dual sim.