**FINAL PROJECT**

PREDICTIVE II

**Akintunde Akinwale**

Title Page 1

Table of Contents 2

Introduction 3

Exploratory plots 5

Discussion of the Plots 6

Model formulation 7

Price classification 9

Gains table 10

The ROC curve 11

The confusion matrix 12

Conclusion 13

**Introduction**

Predictive analytics has become an integral part of statistics and data science which helps in forecasting and predicting alongside determining association rules for quantitative data. This part of statistics / data science incorporates the use of modern statistical methods in modelling data for predictive analytics which uses machine learning methods to derive high-quality performance and meaningful information for all education levels. One approach to this method is using classification, however, this can be subdivided into two major groups (binary and Multiclass classification). This study will be using the multiclass classification method for classifying mobile prices into categories which encompasses low-cost mobile, moderate cost, high cost and very high cost which are the most expensive of the 4 categories.

In recent years, using mobile has become an essential need for the entire human population. This study is aimed at creating a multiclass model that has a high-level accuracy in classifying the price of a mobile phone given the mobile phone specification which includes the software and hardware specification of the mobile phone.

**Data**

The data used for this study is a secondary data which was gotten from a public domain (<https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification>). This data set was divided into two separate files. The train data set and the test data set. The two data splits have 2,000 and 1000 data points respectively.

The data set has about 20 columns which are explained in the table below. All of these formed the predictor variables in the model while price range becomes the target variable for the purpose of this study. In the final multiple linear regression model, these variables were reduced to six.

**Data Dictionary**

|  |  |
| --- | --- |
| **Variable** | **Meaning** |
| battery\_power | Total energy a battery can store in one time measured in mAh |
| clock\_speed | speed at which microprocessor executes instructions |
| dual\_sim | Has dual sim support or not |
| 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 |
| 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 before recharging |
| three\_g | Has 3G or not |
| touch\_screen | Has touch screen or not |
| wifi | Has wifi or not |
| Pc | Has Primary Camera or not |
| Fc | Has Front Camera or not |
| Price range | The price range (Target variable) |

**Exploratory Plots**

A graph of blue rectangular bars

Description automatically generatedA graph of a battery power

Description automatically generated

Fig 1: Showing the distribution of talk time Fig 2: showing the distribution of Battery power.

A graph of a diagram

Description automatically generated with medium confidence

Fig 3: Box plot of RAM vs Price Range (Classes)

A red and blue squares

Description automatically generated

A graph with red squares and blue squares

Description automatically generated

Fig 4: Box plot of Internal memory Vs 4G support Fig 5: 3G and 4G presence

**Discussion of the plots**

**Fig 1:** This figure shows the distribution of the mobile phone talk time. The distribution of the mobile phone talk time is almost uniform. However, the highest talk time is 10 hours. This means that 10hours is the longest time that a single battery charge will last before recharging.

**Fig 2:** This figure shows the distribution of battery power for the data set used in this study. The average battery power falls between 1000 and 1500mAh while the highest battery power is 2000mAh.

**Fig 3:** This shows the box plot of RAM vs Price Range. From the plot, it shows that a mobile phone RAM is a good predictor of its price. This further reveals that mobiles with the least RAM are priced very low and mobiles with a very high RAM have a very high price attached to it.

**Fig 4:** This plot further reveals that mobile phones that support 4G have higher internal memory than mobile phones that do not support 4G networks.

**Fig 5:** This plot shows the clear distinction between mobile phones with 3G and 4G presence. This further shows that mobiles that support 4G also support 3G networks.

**MODEL FORMULATION**

**Multiple linear regression**

The multiple linear regression model is an extension of the simple linear regression model. In simple linear regression, we have one predictor variable that is used to predict the response variable. However, in multiple linear regression, there are several predictor variables that will be used in the model. These predictor variables are assumed to be non-random, full column matrix collected with negligible errors where the error term and the independent variables are independent from each other. One of the assumptions that must not be violated for this type of model is that the model must be linear in terms of regression coefficients.

The equation for the multiple linear regression

y = β0 + β1 X1 + β 2X2 + ... + βj Xj + ε

***Where:***

y = an observed value of the response variable for a particular observation in the population

β0 = the constant term (equivalent to the “y-intercept” in SLR)

βj = the coefficient for the jth explanatory variable

Xj = a value of the jth explanatory variable for a particular observation

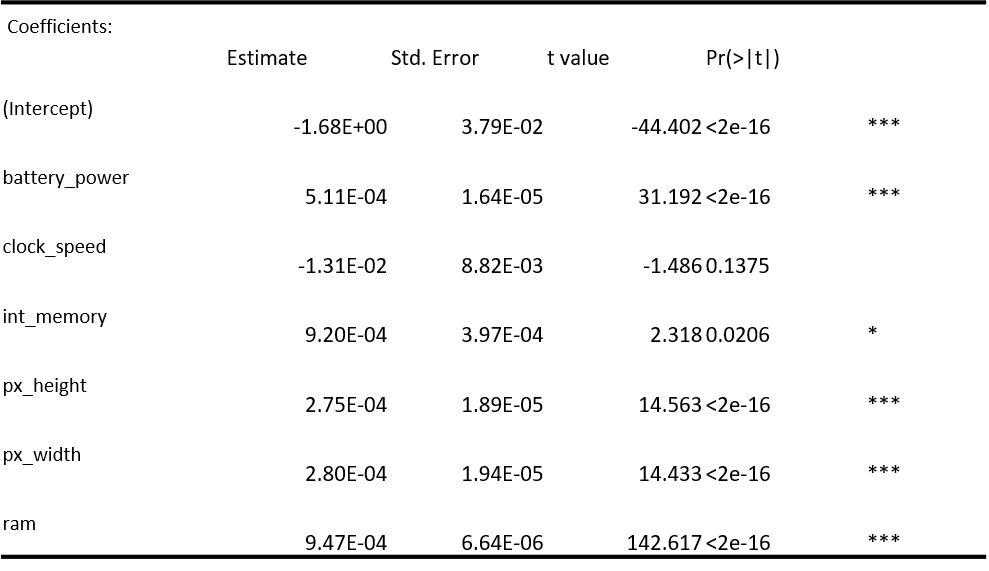
ε = the residual for the observation in the population

Before we arrive at the final model used in this study, several multiple regression models were created. First, all the predictor variables (all columns in the data) were used in the multiple linear regression model. It was discovered that some of these variables are not significant in predicting the price range for a particular mobile device.

The final model that is used in this study, all the predictor variables are significant in predicting the price range for mobile devices except clock speed (speed at which the mobile device microprocessor executes instructions).

**Regression Output:**

The model output was generated using R statistical software which uses battery power, clock speed, internal memory, pixel height, pixel width and RAM as its predictor variables and price range is the response variable.



Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3217 on 1993 degrees of freedom

Multiple R-squared: 0.9175, Adjusted R-squared: 0.9172

F-statistic: 3693 on 6 and 1993 DF, p-value: < 2.2e-16

Price\_range = -1.68E+0 + 5.11E-04\*battery\_power - 1.31E-02\*clock\_speed + 9.20E-04\*int\_memory + 2.75E-04\*px\_height + 2.80E-04\*px\_width + 9.47E-04\*ram

**Discussion:**

The model was able to show the most relevant features that should be considered before buying a mobile phone, and these are the major features that are used to classify mobile prices.

From the model, it reveals that the mobile phone picture quality (px\_height and px\_width) has a significant effect on the mobile price classification. Also, the battery power, RAM and internal memory are also considered as good features to classify the prices.

**PRICE CLASSIFICATION**

**MULTINOMIAL REGRESSION**

**Data Preparation**

As part of the data preparation method, the multiple linear regression model was used for variable selection for the multinomial regression model. This method is more efficient because it allows several model selection techniques. The train data set was further subdivided into two parts for the model training and evaluation purposes. The ratio 70:30 was used for best practices to avoid over fitting of the model.

The multinomial regression used in this study is a statistical method that is used for modelling the relationship between multiple nominal outcomes categories and predictor variables. In this case, the multiple nominal outcomes are the price range which is classified between “0” and “3”. This model is an extension of the widely used binary logistic regression which is used to model the relationship between a binary outcome and its predictor variables.

**Assumptions of the Model**

1. The dependent variable has many levels.

2. Independence of observations: The observations should be independent of each other.

Creating the model used several libraries in R. The set\_engine used for the model formulation was “nnet” and the set\_mode was “classification”.

**Measuring the performance of the model**

In a bid to answer the question “Did it work?” We will be measuring the performance of the models using the Gains Table, the ROC curve, and the Confusion Matrix.

**GAINS TABLE**

A gains table is a tool used in finance to help investors calculate their gains or losses on an investment. The table provides a summary of important information related to the investment, including, total sale proceeds, gain/loss, and percentage gain/loss.

In this context of prediction analysis, the gains table will be used to evaluate the performance of the predicted model. In the multiclass classification problem, this gains table will evaluate the performance of the model that accurately classifies the price of a mobile device given its specification (software or hardware). The gains table evaluates the performance of the model and identifies areas for improvement. Overall, a gains table is a useful tool in prediction analysis for evaluating the accuracy of predictive models and optimizing marketing and sales strategies.

This table shows the gains table for the multiclass classification problem.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Depth**  **Of File** | **N** | **Cume**  **N** | **Mean**  **Resp** | **Cume**  **Mean**  **Resp** | **Cume Pct of Total Resp** | **Lift Index** | **Cume Lift** | **Mean Model Score** |
|  |
|  |
| 27 | 164 | 164 | 3.84 | 3.84 | 41.90% | 153 | 153 | 4 |  |
| 52 | 147 | 311 | 2.8 | 3.35 | 69.40% | 112 | 134 | 3 |  |
| 77 | 149 | 460 | 1.99 | 2.91 | 89.10% | 79 | 116 | 2 |  |
| 100 | 140 | 600 | 1.16 | 2.5 | 100.00% | 47 | 100 | 1 |  |

A graph of a diagram

Description automatically generated with medium confidence

**Interpretation:**

From the table, the model performs best at dept 27. The interpretation for the gains table will be done using the confusion matrix.

**THE ROC CURVE**

In this plot, the True Positive Rate (TPR) is plotted against the False Positive Rate (FPR) at different classification thresholds. The ROC curve visually depicts how well a binary classifier can distinguish between positive and negative classes, and it is used to evaluate the model`s predictive power. A perfect classifier would have an AUC of 1, indicating that it can perfectly distinguish between positive and negative classes. Although, generating the AUC value was almost impossible because the ROCR library in R can only calculate the AUC value for binary response. Hence, we will visually observe the Area Under Curve for the classification.

A graph with different colored lines

Description automatically generated

***This figure shows the ROC curve for the various categories of price classification.***

**Interpretation and Justification**

The ROC curve provides a visual representation of the trade-off between TPR and FPR, allowing for the selection of an optimal classification threshold and the evaluation of the overall performance of a multiclass classification model.

From this figure, it is evident that the price class category (3) has a largest area, followed by class 0, and the class with the least area is class 1. This shows that the model is more efficient in classifying the mobile price in that manner.

Hence, the multiclass model performs better at predicting the most expensive, and the least expensive mobile price while the prediction for categories 1 and 2 are moderate.

**The confusion Matrix**

A confusion matrix is a table that summarizes the performance of a classification model by comparing its predicted labels to the true labels. It displays the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) of the model's predictions.

A grid of squares with numbers

Description automatically generated

The confusion matrix for this study shows that the true positive values for category 3, 2, 1, and 0 and are 137, 95, 90 and 118.

**Category 0:** The False positive values for category 0 that falls under category 1 is 31, under category 2 is 1 and under category 3 is 0 and the False negative for category 0 under category 1 is 21, under category 2 is 1 and under category 3 is 0.

**Category 1:** The False positive values for category 1 that falls under category 2 is 39, under category 3 is 0 and under category 1 is 31 and the False negative for category 1 under category 0 is 21, under category 2 is 27 and under category 3 is 1.

**Category 2:** The False positive values for category 2 that falls under category 1 is 39, under category 0 is 1 and under category 3 is 27 and the False negative for category 2 under category 1 is 27, under category 0 is 1 and under category 3 is 12.

**Category 3:** The False positive values for category 3 that falls under category 2 is 27, under category 1 is 0 and under category 0 is 0 and the False negative for category 3 under category 2 is 12, under category 1 is 1 and under category 0 is 0.

**Conclusion**

From the model performance metrics used in this study, it reveals that the model performance for the model is very good and the predicted categories are very close to the actual categories. Furthermore, the confusion matrix helped us to distinguish the True positive, False negative and False positive values.