**POMS. 6220 DECISION ANALYTICS Course Project**

**Fall 2024**

***From Data to Dollars - Mobile Price Classification***

**A close-up of a phone and a graph

Description automatically generated**

By:

Bhaskar Janani

**Table of Content**

1. Abstract …………………………………………………………………………. 3

2. Keywords ………………………………………………………………………... 3

3. Motivation ………………………………………………………………………. 3

4. Literature Review ………………………………………………………………... 5

5. Predictive Framework …………………………………………………………… 7

6. Results and Analysis ……………………………………………………………. 18

7. Managerial Implication …………………………………………………………. 23

8. Conclusion ………………………………………………………………………. 24

9. Reference ………………………………………………………………………… 25

10. Team Autobiography ……………………………………………………………. 26

1. **Abstract**

The rapid pace of mobile phone technology and market growth makes strategic pricing a critical component of competitive positioning. This study leverages machine learning techniques like Support Vector Machine, Logistic regression, and Decision tree. The logistic regression is used as a standard by which to compare performance, support vector machines (SVM) for managing complicated boundaries and Decision Tree for their ease of use and interpretability - to classify mobile phones into price categories using a dataset with 2000 records and 21 attributes. This study intends to create a strong categorization model by examining how these features affect pricing, offering important insights into feature-driven pricing tactics. Manufacturers and retailers can use the resulting model as a tool to better understand pricing considerations and make well-informed decisions about product positioning and market entry strategies.

This study also explores the influence of key hardware features, such as battery power, camera quality, memory capacity, and screen dimensions, in determining the price category of mobile phones. The model becomes more interpretable by turning some continuous variables into categorical ranges and its conclusions can be used by stakeholders who are non-technical. With an emphasis on feature relevance and optimization for practical use, each machine learning technique was assessed for classification efficiency and accuracy. The findings from this project have the potential to inform both product development and marketing strategies, aiding companies in aligning hardware features with specific pricing tiers to meet consumer expectations and drive profitability. This study seeks to provide valuable insights into how specific features affect pricing, with implications for manufacturers and retailers looking to align product offerings with market demands. By providing more precisely targeted items at competitive prices, the findings also have the potential to increase customer happiness.

1. **Keywords:**

Mobile Price Classification, Predictive Framework, Data-Driven Insights, Support Vector Machine, Decision Tree, Logistic Regression, Hyperparameter, Feature Selection, Error Pruning.

1. **Motivation**

Currently, smartphones have become an intrinsic part of our daily lives, having an essential participation in business, communication, education, and entertainment. The phone market is extensive and extremely rivalry, with multiple manufacturers proving diverse range of devices with boosting distinct features, and cost brackets. Consumer is in dilemma about the phone purchase due to numerous alternatives.

For entrepreneurial ventures and branding strategies, understanding pricing structure and customer inclination is crucial. Cellular manufacturers must be able to spot and categorize the catalysts contributing to the phone’s price and its appeal in different audience demographics. All these factors add up to the need for evidence based prospectives to analyze extensive mobile phone data and leverage this information to predict mobile price categories.

The goal of this study is ‘**mobile price classification**’ , which aims at predicting price range of mobile phones based on numerous functionalities like battery power, bluetooth, clock speed, front camera, touch screen, wi-fi, internal memory, mobile height, mobile width, mobile depth, mobile weight, number of cores, primary camera resolution, pixel resolution height, pixel resolution width, RAM, screen height, screen width, and talk time. The intent of this study lies in mobile phone classification into predefined categories (0-low cost, 1-medium cost, 2- high cost, 3-very high cost) based on mobile features using machine learning algorithms.

Mobile manufacturers can take advantage of predictive models to understand market pricing of their products relative to competitors. By evaluating which features are related to higher price points, they can streamline their product offerings and focus on specific target audiences more effectively. For devising competitive pricing strategies and making data-driven decisions this information will be priceless.

The mobile price classification model can be used by retailers and marketers to project market trends and determine notable features contributing to higher-priced devices. This helps them plan inventory, marketing, and promotional strategies with consumer demand, ensuring that the right products are available at the right price points.

Analysis of this dataset provides prospective of mobile features which strongly impact the mobile prices. Inference of relationship between the distinctive features is advantageous for business to amplify their product sale. The pricing model supports buyers in making astute phone purchasing decisions. The consumer can familiarize with the balance between features and pricing, helping them find the finest device cost-effectively. This model can be integrated with product recommendation systems on e-commerce platforms or comparison websites. The extensive range of mobile phones available in the market complicates the buyer's decision of purchase. By using this model users will find a simple, accessible tool that clarifies and enables more efficient shopping experience.

This project exhibits how machine learning techniques can serve as a tool in retail and pricing strategies, offering a prospect to further develop skills in predive analytics and decision support systems. The aim is to bridge the gap between consumers, producers, and retailers to understand and predict the price range of mobile phones. The project holds the potential to influence decision-making, consumer experience, and build business strategies in the highly competitive mobile phone market.

1. **Literature Review**

Different machine-learning techniques have been studied and evaluated for predictive analysis. Research has been done to identify the price range for Mobile Phone for developing successful pricing for consumers. Mustafa Çetın & Yunus Koç [1] have used datasets from Kaggle.com to build their model. For data preprocessing standard scaling was utilized followed by feature elimination. To include the best features in their model, Mutual Information and ANOVA methods were used because datasets have mostly numerical data and categorical labels where the ANOVA method has the same accuracy as the Mutual Information method was more convenient for this dataset. For both training and test data, hyperparameter optimization was executed. From them, Logistic Regression Classifier, Linear Discriminant Analysis, and SVC models gave good accuracies for both training and testing with low variance. Whereas Random Forest Classifier, Decision Tree Classifier, and k-NN Classifier have high training accuracies while having low test accuracies relatively and have high variance. Overall, for them, it was the SVC classifier that has the highest test accuracy compared to other models. Furthermore, there was no overfitting and underfitting problem in this model.

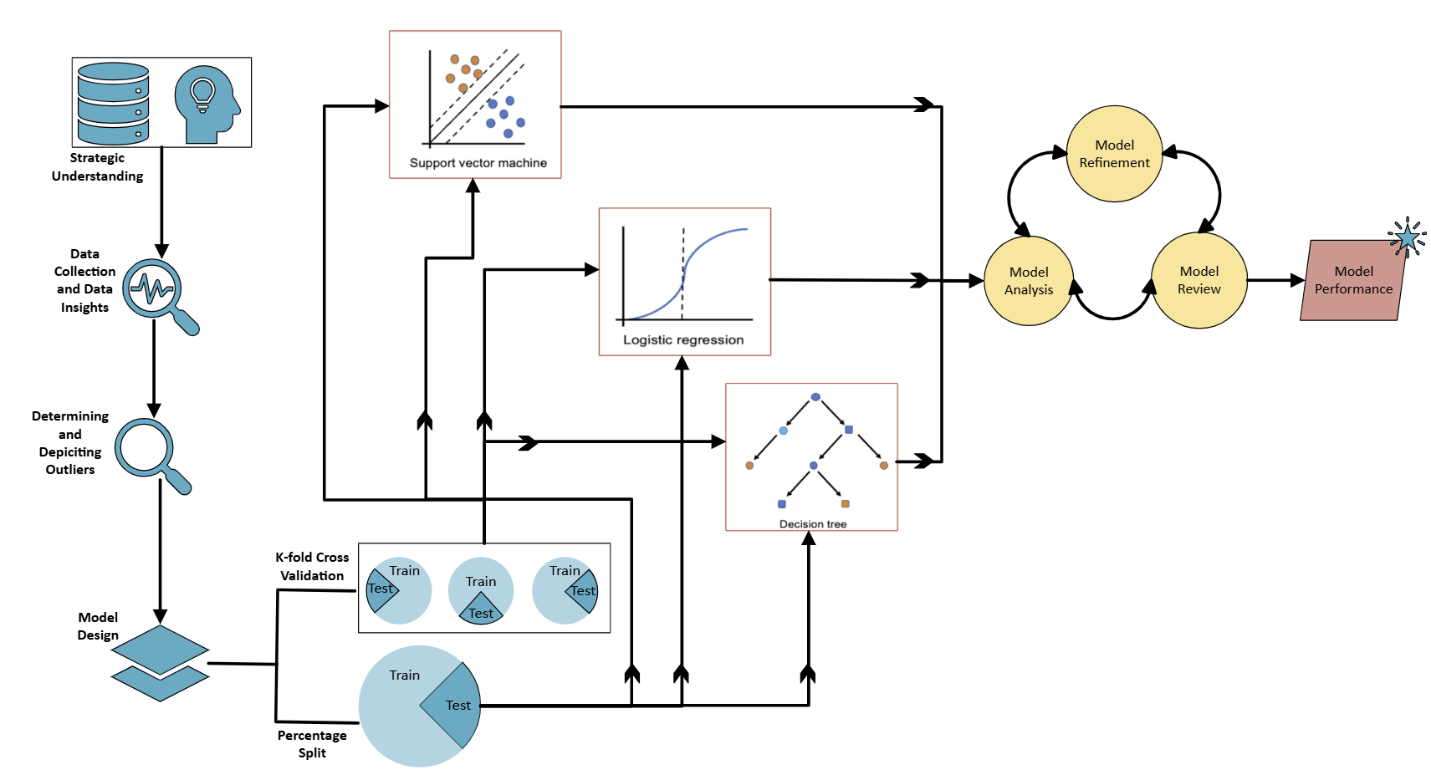
Rwittika, Rakshit, Mahendra, and Pradeep [2], performed a comparative study based on the performance metrics of all the models and classified phone models into four distinct categories namely low, medium, high, and very high cost. In this research, they have used a total of 11 classifiers to predict the class of the mobile price including SVM, k-NN, Naive Bayes, ANN, Xg Boost Classifier, Light Gbm Classifier, Catboost Classifier, Adaptive Boosting Classifier, and for evaluation metrics accuracy, F1 score, sensitivity, and precision were used to analyze the performance of the model. In this work, the accuracy of all the individual classes was mentioned along with the overall accuracy, Precision, sensitivity, and F1 score, and the result was shown through graphs. In the experiment, it was found that linear kernel in SVM was giving more appropriate results. The dataset has shown a linear relationship among its features. XG Boost and Light GBM achieved the same overall accuracy but differed in class-wise performance. The low accuracy of Adaptive Boosting contributes to the model's vulnerability to outliers and noisy data.

Nisha, Avinash, Mehtab, and Vibha [3] have also used machine learning (ML) techniques like SVM, RF, DT, LR, and k-NN to classify and estimate the price range of a mobile phone. In this model again the accuracy of the SVM model was obtained as 98%, whereas the accuracy of the RF model was 88.8%. The Decision tree's accuracy was 80.5 %, compared to 82.6 % for k-NN and 85.5 % for LR. Learning curves were created to illustrate the performance of models concerning the size of the training dataset where SVM maintains a consistently prominent level of performance throughout, indicating its robustness. Menghan Chen [4] predicted the prices of smartphones with features reduction. Without using any feature reduction approaches, Multi-Layer Perceptron (MLP) had a 92.84% accuracy rate. Using principal component analysis (PCA) and Pearson’s correlation as two feature reduction techniques the accuracy suffered. As features were reduced, accuracy went up to 93.22% for the top 15 and 34.06% for the top five. M. Asim and Z. Khan [5] forecasted the price of mobile phones using datasets from GSMArena.com. They gained a maximum accuracy of 78% when WrapperattributEval algorithm was used for feature selection and the Decision tree as a classifier where the combination has selected minimum but most appropriate features. Their other combinations gave 75% (WrapperattributEval algorithm with Naive Bayes), 75% (InfoGainattributEval algorithm with Decision Tree), and 71% (InfoGainattributEval algorithm with Naive Bayes) accuracy. In Amit and Sanjeev ‘s [6] with M5P, they used the divide and conquer strategy to reach decisions and generate if-then rules for performing regression analysis. The M5P algorithm has more accuracy (89%) than the Decision tree (75%) and Tuned Decision tree (85%). Yiwei Chen [7], in his study, achieved a slightly higher accuracy with the k-NN model. The different error rates were calculated by changing the K-value from 1 to 19, where the accuracy of the k-NN model reaches 93.3% When the K value was16, 17, and 19. For Linear Regression, some nonlinear independent variables were deleted from the database. The data was imported into Excel, to perform Linear Regression and calculate its regression coefficient and the accuracy was calculated using Python where the LR model achieved an accuracy of 91.3%. Renuka and Veer [8] have also used Feature selection algorithms and classifiers. This combination achieved maximum accuracy and selected minimum but most appropriate features. It was also observed that adding irrelevant or redundant features to the data set decreases the efficiency of the forward selection method. In contrast, if any important feature is removed from the data set, its efficiency decreases in backward selection.

Like mobile price prediction, ML is used extensively to predict prices in other sectors. Pudaruth [9] investigated the application of ML techniques to predict the price of used cars in Mauritius. The data was collected in less than one-month intervals as time itself could have an appreciable impact on the price of cars in daily newspapers. Using Linear Regression, a remarkably high correlation between the price of a car and the year in which it was manufactured was observed. In the k-NN method prices for Nissan cars turned out to be more consistent than prices for Toyota cars where for Toyota cars, the best value of kwas1 and the performance degrades with increasing values of k while for Nissan cars, the best value of kwas5. Using Naive Bayes and Decision trees the accuracy dangled between 60-70% for different combinations of parameters.

1. **Predictive Framework**

This methodology aims to assess the factors that contribute to various mobile price ranges and classify the mobiles into low cost, medium cost, high cost and very high-cost prices ranges. The observations will help in determining the most effective and accurate method to help customers and business to determine mobile price range based on mobile features. This methodology as shown in figure 1 incorporates strategic understanding, data collection and data insights, determining and depicting outliers, and model design which involves subsequent steps of data split of different models (support vector machines, logistic regression, and decision trees), model analysis, model refinement, model review, and model performance.

Figure 1. Predictive Framework

The selection of SVM, logistic regression, and decision tree is grounded in their wide use across various domains. Their ability to interpret data, robust to outliers, efficiency, versatility, flexibility, and adaptability. Each of these methods is accessed using different percentage splits (60-40, 70-30, 50-50, 30-70, 80-20, 90-10) and cross fold validation (10 and 5-fold cross validation).

**Phase 1: Strategic Understanding**

Mobile price classification dataset holds crucial attributes for predicting mobile price ranges based on mobile features. The mobile phone industry can leverage this classification model to make data-driven decisions, enhance pricing accuracy, market campaigns, and efficient inventory control. The objective of this classification model is to categorize mobile phones into low cost, medium cost, high cost, and very high-cost categories based mobile features like ram, battery power, pixel height., etc. To cater to diverse customers strategic pricing in the mobile industry is required. By data analysis companies can aim at determining phone characteristics contributing towards price segments. Businesses can optimize their pricing strategy and make informed decisions on discounts, promotions, and plan product bundles. This dataset also reveals rising demand for storage space in phones, battery life, and the need for advanced cameras. Insights into feature-based mobile price classification will help the business in understanding customer purchasing motivation. This classification model will also help the customer in making informed decisions about phone purchase.

**Phase 2: Data Collection and Data Insights**

The data is sourced from Kaggle in a csv file. Thorough examination of the dataset was conducted by reviewing the data, independently examining individual columns, and identifying the outliers, missing values, and relation within different attributes which could impact the model analysis and prediction result.

**Variables used in this Study with Type and Explanation**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Variable Type** | **Explanation** |
| batry\_pwr | Numeric | Battery power of the mobile (in mAh). |
| bluetooth | Binary (0, 1) | Indicates if the mobile has Bluetooth (0- No, 1 - Yes). |
| clck\_spd | Numeric | Clock speed of the mobile processor (in GHz). |
| dual\_sim | Binary (0, 1) | Indicates if the mobile has dual SIM support (0 for No, 1 for Yes). |
| front\_cam | Numeric | Front camera resolution (in megapixels). |
| four\_g | Binary (0, 1) | Indicates if the mobile supports 4G connectivity (0 – No, 1 for Yes,). |
| intrnl\_mem | Numeric | Internal memory/storage of the mobile (in GB). |
| mbl\_depth | Numeric | Depth or thickness of the mobile (in cm). |
| mbl\_wght | Numeric | Weight of the mobile (in grams). |
| no\_cores | Numeric | Number of cores in the mobile's processor. |
| prm\_cam\_mp | Numeric | Primary (rear) camera resolution (in megapixels). |
| px\_res\_ht | Numeric | Pixel resolution height of the screen. |
| px\_res\_wdt | Numeric | Pixel resolution width of the screen. |
| ram\_mb | Numeric | Random Access Memory (RAM) of the mobile (in MB). |
| scrn\_ht | Numeric | Height of the mobile screen (in cm). |
| scrn\_wdt | Numeric | Width of the mobile screen (in cm). |
| talk\_time | Numeric | Maximum talk time (in hours) supported by the mobile on a full charge. |
| three\_g | Binary (0, 1) | Indicates if the mobile supports 3G connectivity (0 – No, 1 - Yes). |
| touch\_scrn | Binary (0, 1) | Indicates if the mobile has a touchscreen (0 - No, 1 -Yes). |
| wifi | Binary (0, 1) | Indicates if the mobile has WiFi connectivity (0 - No, 1 - Yes). |
| price\_range | Categorical | Price range category of the mobile (0 for low, 1 for medium, 2 for high, 3 for very high). |

**Distribution Of Different Price Ranges:** The pie chart shown in figure 2, provides streamlined insights of distribution of mobile phones across different price ranges. Each slice of pie represents the percentage of the dataset that belongs to a particular price range. The pie chart concludes that low (0), medium (1), high (2), and very high cost (3) categories are distributed evenly across the entire data set.

**A pie chart with different cost levels

Description automatically generated**

Figure 2. Price Range Distribution

**Correlation Heatmap:** Figure 3 displays the relationship between distinctive features in the dataset. By picturing a correlation heatmap one can effortlessly depict the patterns and relations in the dataset. The correlation Heatmap interpret that the following features have positive correlation, for example as ram\_mb increases price\_range increases.

1. ram\_mb and target variable price\_range.
2. three\_g and four\_g
3. prm\_cam\_mp and front\_cam
4. px\_res\_wdt and px\_res\_ht
5. scrn\_wdt and scrn\_ht

**A blue and red graph

Description automatically generated**Figure 3. Correlation Heatmap

Comprehending the distribution of ram\_mb and batry\_pwr against price\_range: From figure 4 it can be recognized that as the ram\_mb and batry\_pwr increase the phone price increases.

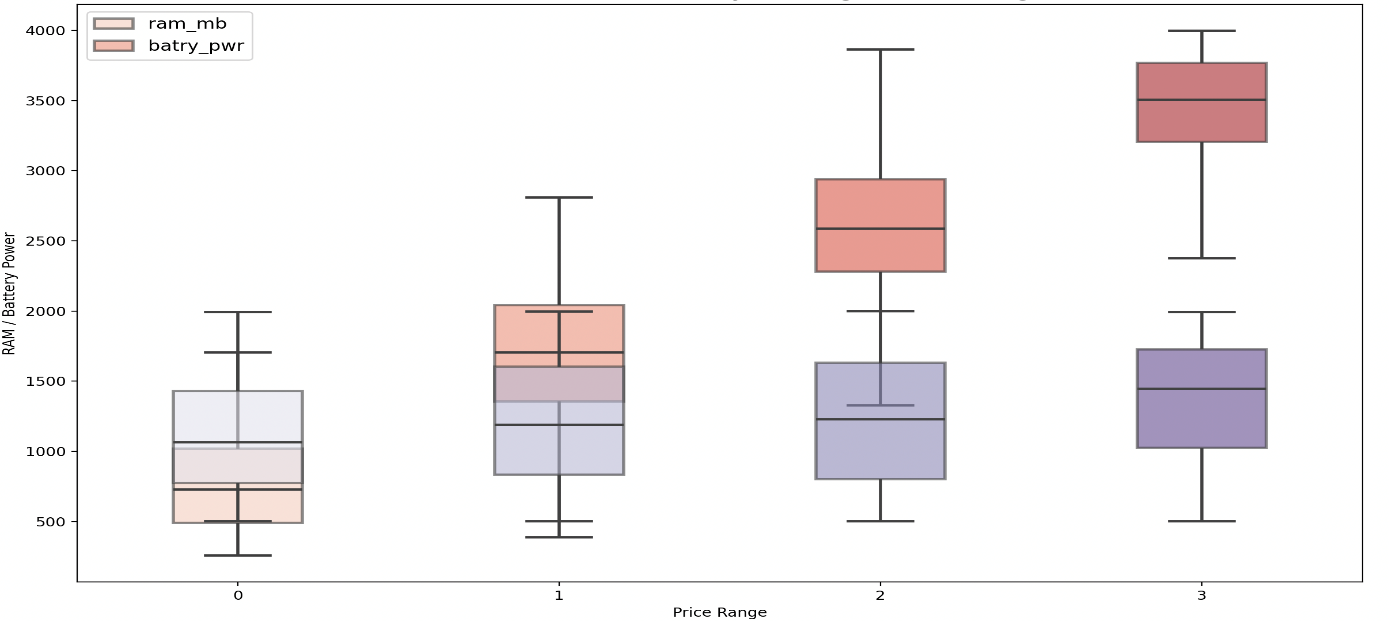


Figure 4. Distribution of RAM and Battery Power against Price Range

Comprehending the distribution of px\_res\_wdt and px\_res\_ht against price\_range: Figure 5 depicts as the px\_res\_wdt and px\_res\_ht increases the price range increases but it remains persistent across medium and high-cost phone prices.

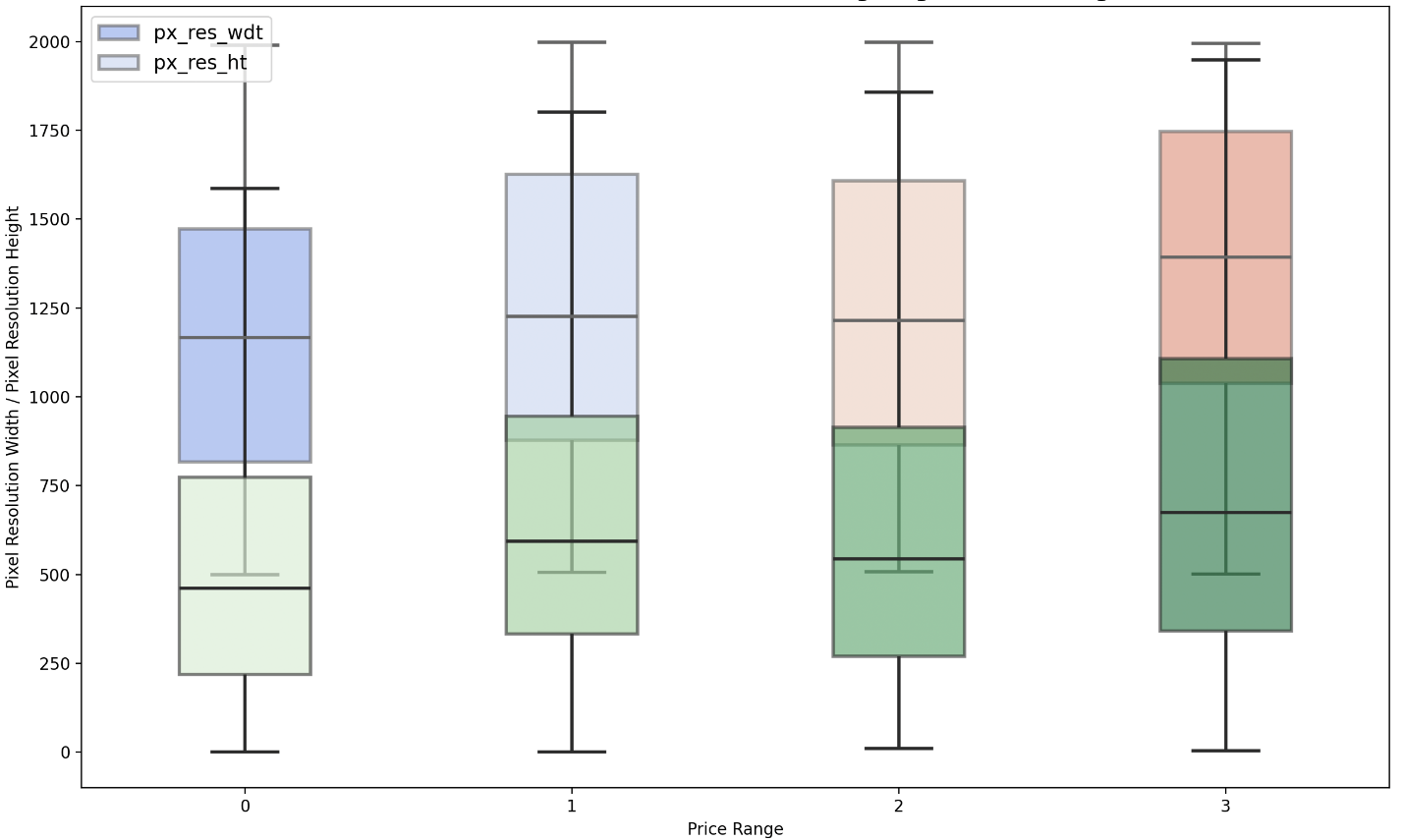


Figure 5. Distribution of Pixel Resolution Width and Height against Price Range

**Phase 3: Determining and Depicting Outliers:**

The missing values were checked using Weka's ‘Missing’ feature and duplicates were verified using excel. A gripping arrangement was observed where records had primary camera megapixel of 0 but still had pixel resolution width and height as shown in figure 6 and 7. These records illustrate outliers as it is uncommon for a mobile to have pixel resolution width and pixel resolution height without the primary camera. There were 101 such records in the dataset which could skew the analysis and hence they were removed from the dataset manually.

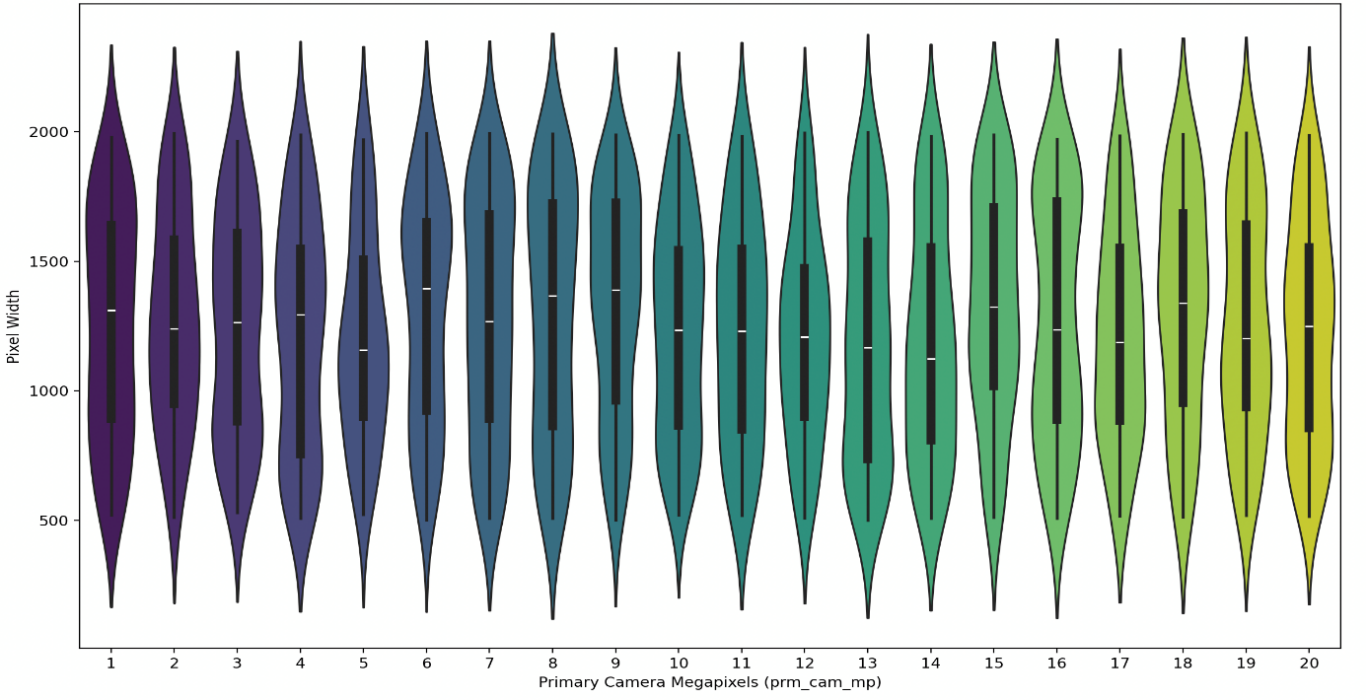


Figure . Distribution of Pixel Width by Camera Megapixels

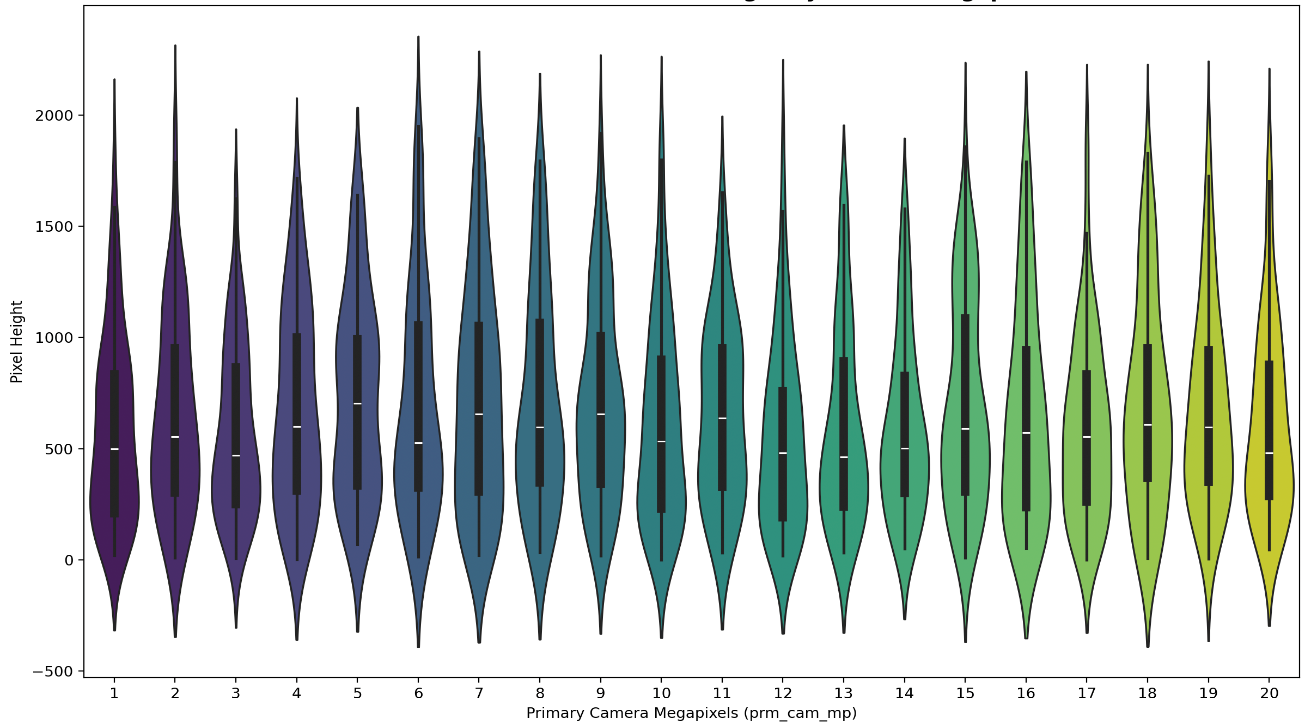
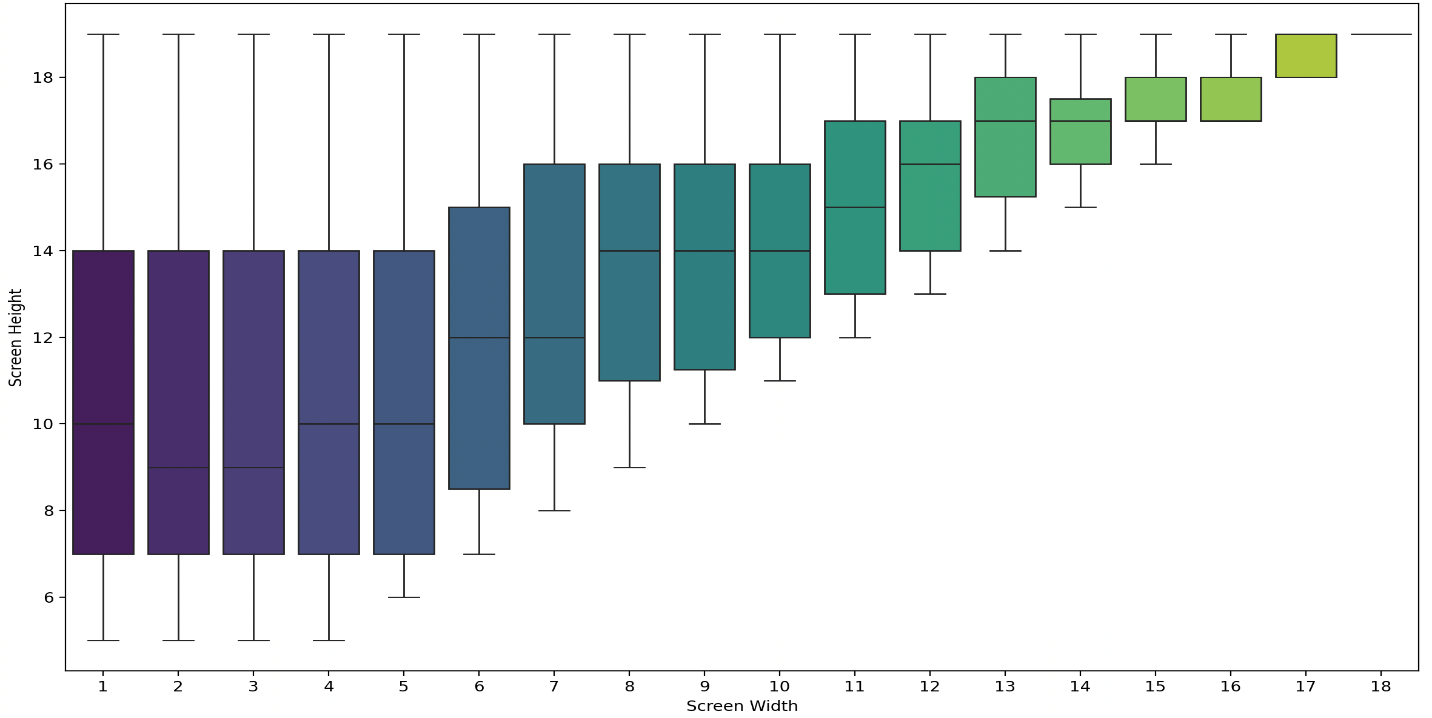


Figure 7. Distribution of Pixel Resolution Height by Camera Megapixels

The data holds records with screen width of 0cm but still has screen height as shown in figure 8. Screen dimensions usually have screen width and screen height, but this dataset had 180 records with 0 screen width with screen height indicating the abnormality in the data. As these records can skew the data analysis and could lead to misinterpretation of data they were manually removed.

Figure 8. Screen Dimensions Distribution

**Phase 4: Model Design**

**Phase 4.1: K-fold cross validation**

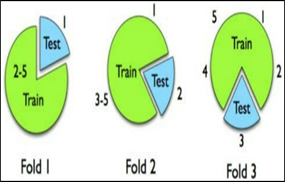
K-Fold CV is a common shorthand used in data science and machine learning. The dataset is divided into K equally sized folds**.** Each fold is used once as a test set, while the remaining K-1 folds form the training set. This process is repeated K times, ensuring every data point is evaluated exactly once. The final performance is averaged across all folds to provide a robust estimate. It helps prevent overfitting and ensures the model's evaluation is unbiased. Additionally, K-fold cross-validation is highly effective for fine-tuning hyperparameters, leading to improved model performance.

Figure 9. K-fold

**Phase 4.2: Percentage split**

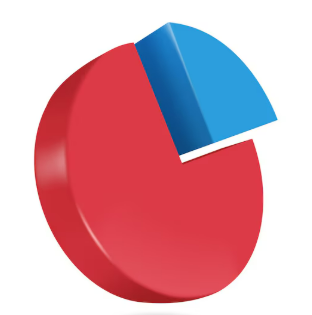
Percentage Split is also referred to as Train-Test Split, emphasizing the division into training and testing datasets. The dataset is split into two parts: a training set and a test set based on a predefined percentage (e.g., 70% training, 30% testing). The model is trained on the training set and evaluated on the test set. It is simpler and faster than K-fold but may be biased if the split does not represent the dataset's overall characteristics. Commonly used in scenarios where data volume is large, and repeated testing is not necessary. Ideal for quick evaluations but less robust than K-fold cross-validation.

Figure 10. Percentage Split

**Model 1: Support Vector Machine (SVM)**

Finding the optimal hyperplane that separates data into distinct classes is the goal of the supervised learning model SVM, which is used for regression and classification. For linearly separable data, it maximizes the margin between classes, while for non-linear data, it uses kernels to map data to higher dimensions. It is effective in high-dimensional spaces and robust against overfitting with proper regularization. SVMs can be computationally intensive, especially with large datasets.

**Model 2: Logistic Regression**

A statistical technique for binary or multi-class classification problems is logistic regression. It predicts the probability of an outcome by fitting data to a logistic curve using a sigmoid function. The model estimates the relationship between input features and the log-odds of the target class. It is computationally efficient and interpretable but assumes linear separability of data. Logistic regression is less effective for complex, non-linear relationships unless combined with feature engineering.

**Model 3: Decision tree**

Decision trees are intuitive and versatile models used for classification and regression. It creates a tree-like structure by dividing data into branches according to requirements. Each internal node represents a decision based on a feature, branches represent outcomes, and leaves represent final predictions. Decision trees are easy to interpret but can overfit if not pruned. They can effectively manage numerical and categorical data.

**Feature Importance Analysis:**

Feature importance analysis as shown in figure 11 using python revealed ram\_mb, px\_res\_wdt, batry\_pwr, and px\_res\_ht are the only important variables that contribute towards price range classification.

**A white background with blue lines

Description automatically generated**

Figure 11. Feature Importance

The performance of decision trees, logistic regression, and support vector machines with all variables and with featured variables is as shown in figure 12. The results indicate that the performance across decision tree and logistic regression models with all variables and with featured variables is identical whereas it shows minimal variations in SVM model, as the difference in accuracy was very minimal **t**o ensure explainability, abstain data overfitting, and to avoid complex model nature, all the models were designed with only featured variables.

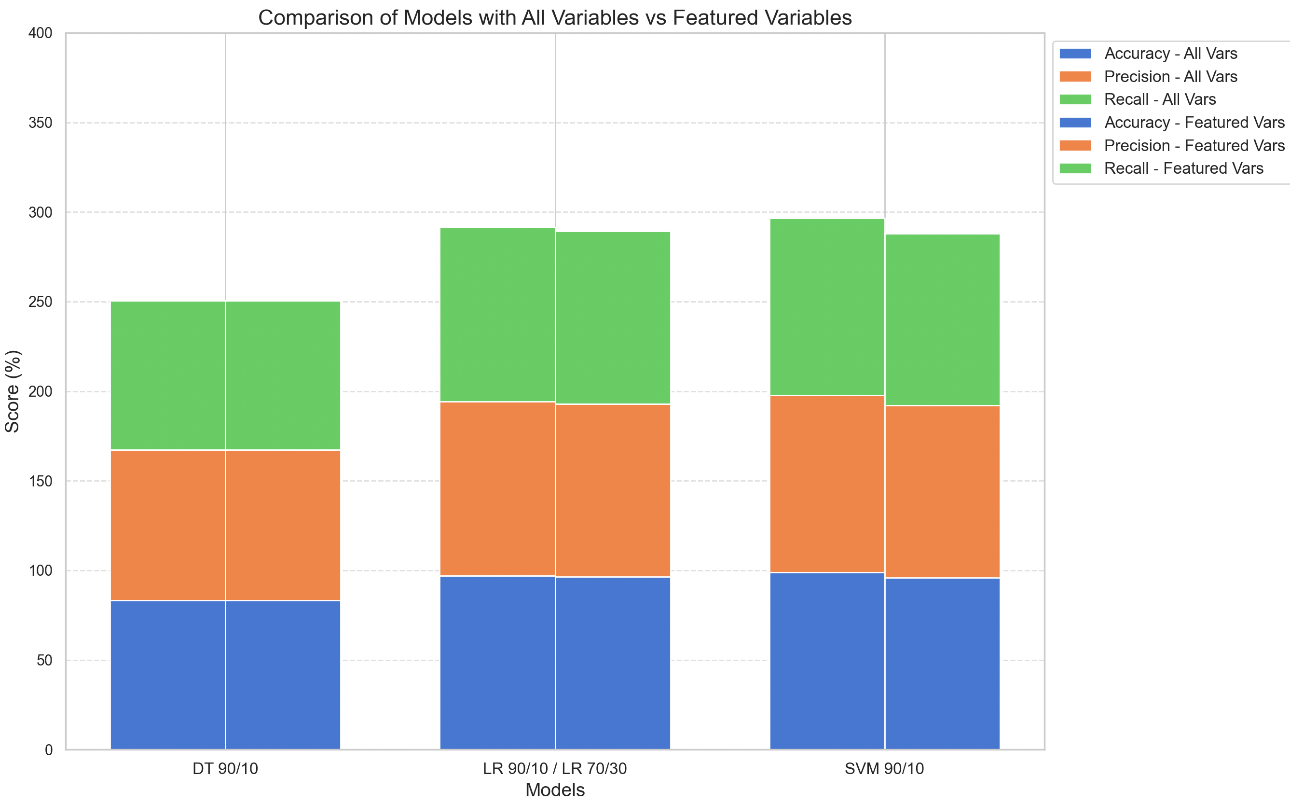


Figure 12. Model Comparison with All Variables vs Featured Variables

1. **Results And Analysis**

**Model Analysis, Model Refinement, Model Review, and Model Performance**

The last phase of the framework is to analyze, refine, and review the model's performance. The measures used for evaluating models' performance are shown in Table 1.

|  |  |
| --- | --- |
| **Table 1: Model Effectiveness Measure** | |
| Accuracy: Measures the proportion of correctly classified instances out of all instances. | Where T is the number of correctly classified records |
| Precision: Measures the proportion of correctly predicted positive cases out of all predicted positive cases | Where T is the number of correctly classified records and e is the number of error values |
| Sensitivity: Also known as recall, measures the proportion of actual positive cases that are correctly identified | Where T is the number of correctly classified records and e is the number of error values |
| Specificity: Is the ability of a model to correctly identify true negative rate. | Where T –1 is the number of unpredicted values and e is the number of error values |

Considering featured variables, the performance of three models with different percentage splits and cross fold validation has been computed in Table 2.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 2: Model Performance with Different Percentage Split and K-Folds** | | | | | | |
| **Support Vector Machine** | | | | | | |
| **Percentage split**  **K-Folds** | **Accuracy** | **misclassified Instances** | **Precision** | **Sensitivity** | **F-Measure** |
| 90/10 | 96.53% | 3.47% | 0.968 | 0.965 | 0.966 | |
| 80/20 | 95.09% | 4.91% | 0.953 | 0.951 | 0.951 | |
| 70/30 | 96.35% | 3.65% | 0.965 | 0.963 | 0.964 | |
| 60/40 | 95.96% | 4.04% | 0.96 | 0.96 | 0.96 | |
| 50/50 | 95.27% | 4.73% | 0.953 | 0.953 | 0.953 | |
| 30/70 | 94.97% | 5.03% | 0.952 | 0.95 | 0.95 | |
| 5-Fold | 95.84% | 4.16% | 0.958 | 0.958 | 0.958 | |
| 10-Fold | 95.61% | 4.39% | 0.956 | 0.956 | 0.956 | |
| **Logistic Regression** | | | | | | |
| **Percentage split** | **Accuracy** | **misclassified Instances** | **Precision** | **Sensitivity** | **F-Measure** | |
| 90/10 | 95.38% | 4.62% | 0.957 | 0.954 | 0.954 | |
| 80/20 | 96.24% | 3.76% | 0.964 | 0.962 | 0.963 | |
| 70/30 | 96.35% | 3.65% | 0.966 | 0.963 | 0.964 | |
| 60/40 | 95.96% | 4.04% | 0.961 | 0.960 | 0.960 | |
| 50/50 | 94.80% | 5.20% | 0.950 | 0.948 | 0.948 | |
| 30/70 | 95.63% | 4.37% | 0.958 | 0.956 | 0.957 | |
| 5-Fold | 95.79% | 4.21% | 0.958 | 0.958 | 0.958 | |
| 10-Fold | 95.67% | 4.33% | 0.957 | 0.957 | 0.957 | |
| **Decision Tree** | | | | | | |
| **Percentage split** | **Accuracy** | **misclassified Instances** | **Precision** | **Sensitivity** | **F-Measure** | |
| 90/10 | 83.24% | 16.76% | 0.841 | 0.832 | 0.833 | |
| 80/20 | 81.21% | 18.79% | 0.829 | 0.812 | 0.814 | |
| 70/30 | 77.31% | 22.69% | 0.78 | 0.773 | 0.774 | |
| 60/40 | 74.31% | 25.69% | 0.761 | 0.743 | 0.745 | |
| 50/50 | 70.67% | 29.33% | 0.72 | 0.707 | 0.702 | |
| 30/70 | 74.59% | 25.41% | 0.757 | 0.746 | 0.749 | |
| 5-Fold | 78.35% | 21.65% | 0.78 | 0.783 | 0.777 | |
| 10-Fold | 80.20% | 19.80% | 0.798 | 0.802 | 0.797 | |

Considering the models with highest accuracy, the precision, sensitivity, and specificity per class is tabulated in Table 3.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 3: Confusion Matrix** | | | | | | | |
| Decision Trees (90% training, 10% testing) | | | | | | | |
| **Classified as ->** | **Medium Cost** | **High Cost** | **Very High Cost** | **Low Cost** | **Precision** | **Sensitivity** | **Specificity** |
| **Medium Cost** | 31 | 10 | 0 | 4 | 0.816 | 0.689 | 0.95 |
| **High Cost** | 6 | 32 | 2 | 0 | 0.667 | 0.800 | 0.88 |
| **Very High Cost** | 0 | 6 | 38 | 0 | 0.950 | 0.864 | 0.98 |
| **Low Cost** | 1 | 0 | 0 | 43 | 0.915 | 0.977 | 0.97 |
| Logistic Regression (70% training, 30% testing) | | | | | | | |
| **Classified as ->** | **Medium Cost** | **High Cost** | **Very High Cost** | **Low Cost** | **Precision** | **Sensitivity** | **Specificity** |
| **Medium Cost** | 137 | 4 | 0 | 0 | 0.958 | 0.972 | 0.98 |
| **High Cost** | 0 | 115 | 1 | 0 | 0.906 | 0.991 | 0.97 |
| **Very High Cost** | 0 | 8 | 130 | 0 | 0.992 | 0.942 | 1.00 |
| **Low Cost** | 6 | 0 | 0 | 119 | 1.000 | 0.952 | 1.00 |
| Support Vector Machines (90% training, 10% testing) | | | | | | | |
| **Classified as ->** | **Medium Cost** | **High Cost** | **Very High Cost** | **Low Cost** | **Precision** | **Sensitivity** | **Specificity** |
| **Medium Cost** | 44 | 1 | 0 | 0 | 0.957 | 0.978 | 0.98 |
| **High Cost** | 0 | 40 | 0 | 0 | 0.909 | 1.000 | 0.97 |
| **Very High Cost** | 0 | 3 | 41 | 0 | 1.000 | 0.932 | 1.00 |
| **Low Cost** | 2 | 0 | 0 | 42 | 1.000 | 0.955 | 1.00 |

Sequential Minimal Optimization (SMO) classifier was used in developing support vector machines with 90% training and 10% testing data, the complexity parameter was set to 10 to control the trade-off between smooth decision boundary and some misclassification of training instances. To set high cost for misclassification polynomial kernel with degree = 1 was used. The model gave accuracy of 96.53%. The negative weights of ram\_mb (-4.3701), batry\_pwr (-1.0494), px\_res\_wdt (-0.6388), and px\_res\_ht (-0.8977) indicate an increase or decrease in these variable values tend to push the class towards a very high or low-cost price range.

In developing logistic regression model the data was split into 70% training and 30% testing, with regularization applied to prevent overfitting of data 96.35% of accuracy was achieved. With low-cost class as the reference the positive coefficients of ram\_mb, batry\_pwr, px\_res\_wdt, and px\_res\_ht indicate the mobile price failing in medium, high, or very cost range as in Figure 13.

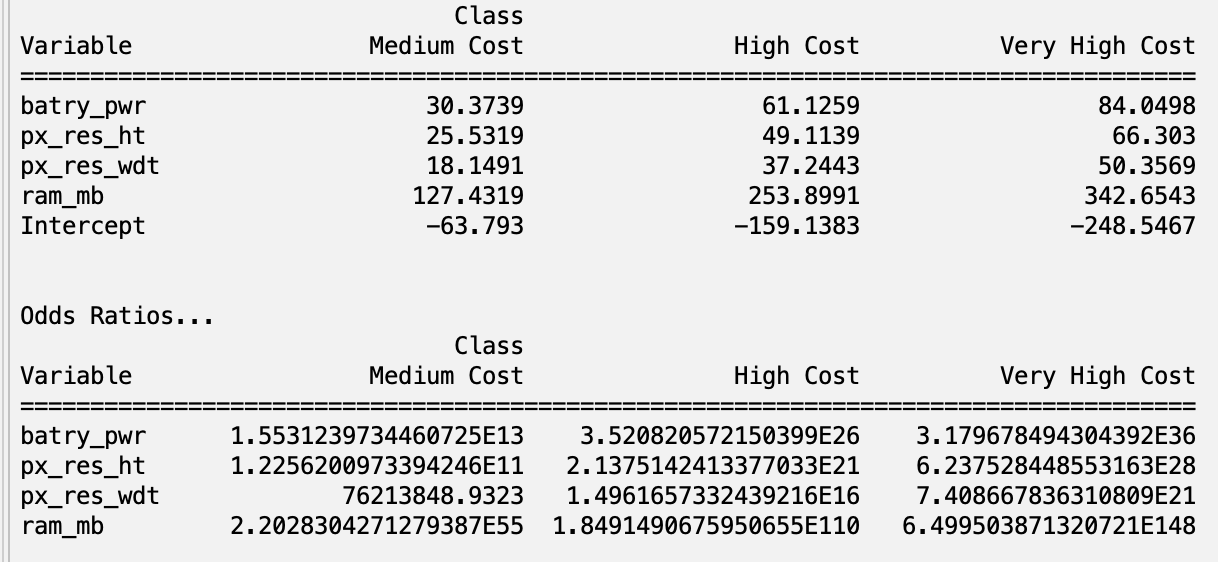
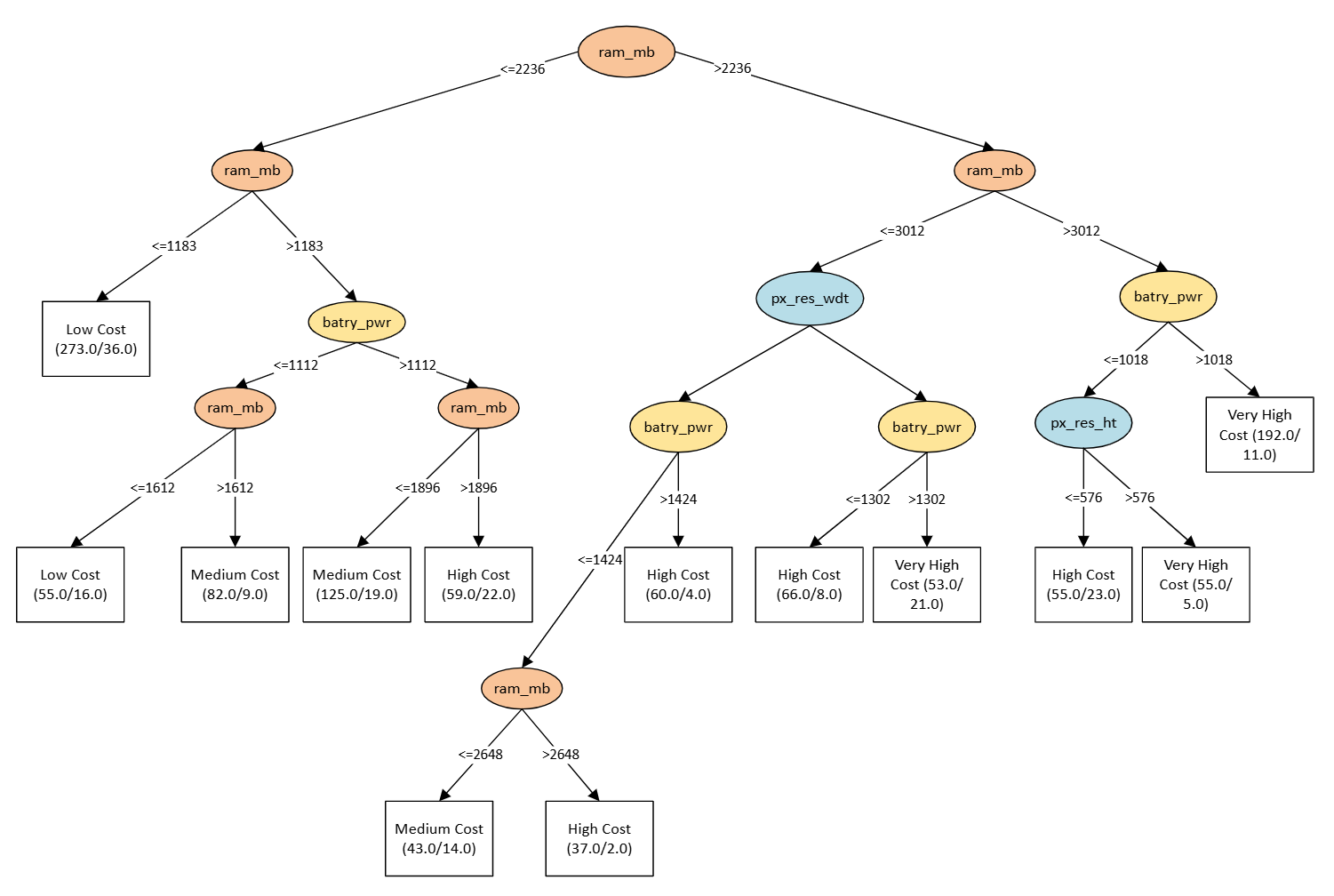


Figure 13. Logistic Regression Variable Coefficient

Also, the high odds ratio of ram\_mb and batry\_pwr also indicates the strong influence of these variables on price range prediction. Logistic regression has slightly balanced matrix especially for medium cost whereas SVM has better precision for all classes especially for low and very high cost. Though logistic regression and SVM gave better accuracy their interpretability was not clear hence the decision tree was designed to get a better picture of the influence of the variables on the price prediction.

The decision tree with featured variables and fine-grained splits (minNumObj=2) gave accuracy of 86.7% but even nodes containing just two data points led to the creation of complex tree with insignificant patterns. After increasing the number of objects to thirty the tree algorithm required at least 30 instances before it could split further so this prevented overfitting of data and simplified tree. Also enabling reduced error pruning to ‘True’ in Weka reduced the number of branches which corresponds to overfitting the data. This resulted in more generalized splits to make a compact decision tree. While the tree is simple and clearer, it reduced the model's accuracy to 83.24% indicating the model does not capture the details of the data well. But reduced accuracy can be treated as a balancing act of improved interpretation of the decision tree. From figure 14 it is evident that increasing values of ram\_mb, battery, px\_res\_wdt, and px\_res\_ht corresponds to higher price range.

Figure 14. Decision Tree

1. **Managerial Implications**

Price is a key aspect of any business. It plays a significant role in value creation. It is important to capture value through the right pricing strategy. Currently the phone market is highly competitive where multiple brands can cater to every consumer requirement. All these existing mobile manufacturing companies already have a reference price in the market. With this application of classification of mobile price, it will help the brand determine their price for their product depending on the mobile phone's features and components. They can fit their device selling price in price elasticity by categorizing the mobile phones under different class range. They will be able to prioritize the prominent features of their device that play a critical role in determining the overall performance of mobile devices. This analysis enables the decision makers to draw relations among the features and price. With the application of different methods, the managers can decide which features enhance the device and lead to higher prices and which features are not significant players in price determination. With this analysis, they can also develop a price-quality relationship and build a strategy to rightly place their product. With this knowledge of price classification, the sales team can also use predictive analysis to forecast their sales and estimate their revenue generation. The mobile phones which fall under extremely high cost can be marketed as their premium product by highlighting its best features and on the other hand, mobile phones from low cost or medium cost category can be marketed as affordable option, offering value for money with essential features. This analysis and its application can be extended to other fields such as car price prediction, property price prediction, oil price prediction, gold rate prediction. This study enables us to use this application in the real world in different industry sectors.

1. **Conclusion**

Mobile phones are available in different price ranges for all budget requirements. But it is challenging to make informed purchasing decisions as a consumer and for the business it is strenuous to target the segment of customers. The other challenge with the wide price range of phones is surviving in the competitive landscape of the mobile world. The use of data driven decision making tools to accurately classify the phones based on their features is a bliss to categorize product, pricing strategy, customer segmentation, and to stay relevant in the competitive market. Consumers can utilize these tools to make informed purchasing decisions. The results obtained from this innovative predictive framework reveal that mobile price data can be analyzed and interpreted using data mining techniques to obtain useful information to solve mobile commerce industry problems.

The data mining models used in this study offer promising approach for understanding the factors influencing mobile pricing and classifying the phones into different price ranges. From all the three different algorithms it is evident that each model has its own strength in depicting the price range based on mobile features. It is transparent from Table 2 and Table 3 that out of all three models support vector machines (SVM) has the highest accuracy of 96.53% with no misclassification of high-cost instances(sensitivity=1.00). With high average precision and sensitivity for low, medium, high, and very high-cost instances SVM should be chosen to predict the price range, for proper inventory management, customer segmentation, promotion, recommendations, and for resource allocation. Support vector machines should be the choice with the multiclass, complex class boundaries, when accuracy is most important and the need is to capture non-linear relationship between ram\_mb, batry\_pwr, and other variables. But on the other hand, logistic regression also performs well with accuracy of 96.35% and in comparison, with SVM high precision for medium cost signifies its ability to correctly classify the instances across the medium class. Logistic regression is simple to interpret and from its coefficients it can be sighted that ram\_mb, batry\_pwr, px\_res\_wdt, and px\_res\_ht indicate the strong predicting influence on price\_range failing in medium, high, or very cost range. The same can be interpreted though decision trees but it has the lowest accuracy of 83.24 % with more misclassifying instances. The interpretability of decision trees gives clear understanding that the instances with greater ram, battery power, pixel resolution height and width pushes the phone range into high or very high cost. On the other hand, the lower value of these variables contributes to the Low and Medium phone cost. Decision tree faces challenges with classifying medium and high-cost instances. But its higher sensitivity value of 0.977 shows its ability to correctly classify the low-cost phones.

Despite the models' ability to predict mobile phone prices further analysis is required in tuning the parameters and reducing misclassification of instances. Adding a few additional parameters display features, battery life, internal storage, and launch year., etc. which could enhance the model's performance. Overall, these models augmented with data driven predictive framework and feature important analysis to classify and identify the instances into different price range based on the mobile features can be the decision-making tool for business and can also be used for getting insights into mobile features, customer segmentation, making marketing strategies, and focusing on product development.

1. **Reference**

[1] Mobile Phone Price Class Prediction Using Different Classification Algorithms with Feature Selection and Parameter Optimization, Mustafa Çetın; Yunus Koç

[2] Comparative Analysis of Mobile Price Classification Using Feature Engineering Techniques, Rwittika Pramanik; Rakshit Agrawal; Mahendra Kumar Gourisaria; Pradeep Kumar Singh

[3] Classification of Mobile Price Using Machine Learning- NishaSunariya1, Avinash Singh, Mehtab Alam, and Vibha Gaur, Department of Computer Science, University of Delhi, New Delhi, India

[4] Mobile Phone Price Prediction with Feature Reduction, Menghan Chen \*Department of Computing Science, University of Alberta, Canada

[5] Mobile Price Class prediction using Machine Learning Techniques, International Journal of Computer Applications (0975 – 8887) Volume 179 – No.29, March 2018, Muhammad Asim, Zafar Khan

[6] Performance Analysis of Decision Tree Models and M5P Models for Mobile Phone Price Prediction, Amit Kumar Bishnoi; Sanjeev Kumar Mandal

[7] Prediction of Different Types of Mobile Phone Prices based on Machine Learning Models, Yiwei Chen from Zhejiang University of Technology, China

[8] Online Mobile Price Prediction using Machine Learning, Turkish Online Journal of Qualitative Inquiry (TOJQI), Miss. A. Renuka, Mr. Veer Sudheer Goud, Holy Mary Institute of Technology & Science, Hyderabad, India.

[9] Predicting the Price of Used Cars using Machine Learning Techniques, Sameerchand Pudaruth, Computer Science and Engineering Department, University of Mauritius, MAURITIUS.

[10] Platt J (1998) Fast training of support vector machines using sequential minimal optimization. Schölkopf B, Burges C, Smola AJ, eds. Advances in Kernel Methods – Support Vector Learning (MIT Press, Cambridge, MA), 185-209.

[11] Patel N, Upadhyay S. Study of various decision tree pruning methods with their empirical comparison in WEKA. Int J Comp Appl. 60(12):20–25.

[12] Berry MJA, Linoff G. Mastering Data Mining: The Art and Science of Customer Relationship Management. New York: John Wiley & Sons, Inc; 1999.

[13] MeSH B. (2005) Bethesda: National Library of Medicine. Available at http://www.nlm.nih.gov/mesh/MBrowser.html. Retrieved April 2, 2007.

[14] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf. “Support vector machines.” Intelligent Systems and their Applications, IEEE, 13(4), pp. 18–28, 1998.

1. **Team Autobiography**

In 2017 Nainul Arab earned bachelor's in electrical and Electronics Engineering from PES college of Engineering, India. With 6.5 years of experience as a data engineer Nainul has demonstrated her ability to analyze and develop data driven solutions. She has a strong record of planning,

A person smiling at camera

Description automatically generatedJanani Bhaskar completed a bachelor's degree in commerce with a specialization in Accounts in 2021 and a master's degree in human resource management in 2023. Currently she is pursuing an MS in Business Analytics at the University of Massachusetts Lowell, focusing on developing advanced data analysis and decision-making skills. During her master's in human resource management, she has actively contributed to various initiatives, including sports leadership, employee engagement activities, and club volunteering, showcasing a strong ability to collaborate and lead. She is passionate about leveraging her diverse experiences to excel in data analytics and business intelligence roles. She combines her HR experience with a passion for analytics to deliver insightful solutions to business challenges and to bridge the gap between people management and data-driven solutions