Mobile price classification

Vishal Agarwal 210962148
Joy Podder 210962184
CSE AIML
MIT Manipal

***Abstract*—This paper classifies a mobile into various price segments based on its specifications such as RAM, battery, pixel height and width, processor clock speed, a total of 20 such features are present in the dataset used. After the data is preprocessed and normalized, we have used feature engineering techniques to calculate the mRMR values and choose features that play an important role in determining a mobile's price segment. These features are then passed through classifiers, namely KNN classifier, logistic regression classifier, decision tree classifier and random forest classifier. The results obtained are then visualized.**

Keywords—preprocessed, normalized, feature engineering, classifiers, visualized.

I. Introduction

In today's constantly changing and extremely competitive world manufacturers often face the challenge of launching their products in the right price segment in the market. A product's price determines its target audience and is a very important factor in deciding whether the product is a success or not. This report deals with the classification of mobile devices into distinct price segments. This classification is not only crucial for setting competitive prices but also for understanding customer preferences and staying ahead of the competition.

Our model aims to help in this classification by leveraging feature engineering techniques and classification algorithms. By comparing results from 4 classifiers our model serves as a predictive tool that can categorize devices into specific price brackets.

This report provides an overview of the methodology, preprocessing techniques, classifiers, and evaluation metrics used in the creation of this model. The significance of this project lies not only in its potential to simplify decision-making processes for mobile manufacturers but also in contributing to the broader perspective of the intersection of artificial intelligence and the mobile industry. Throughout the report the steps taken to construct the model will be elucidated and validated, and finally its results will be presented.

II. Literature Review

A. Mobile price class prediction with feature selection and parameter optimization

As seen in [1] a similar dataset is used as this model to classify which price segment a mobile phone belongs to. The paper uses Random Forest, Logistic Regression, Decision Tree, Linear Discriminant Analysis, K-Nearest Neighbour Classifier and SVC methods to achieve results. It employs the ANOVA f-test method to select features that are significant. The model finally selected is SVC classifier.

B. XGBoost model for mobile price classification

In [2] DBO and Xgboost is used to classify mobile prices In the PCA dimensionality reduction phase, a feature filtering strategy has been devised by considering feature importance., The parameters of XGBoost are selected through DBO to build the model. The experimental results state that the DBO and XGBoost model performs 95.5% better than a traditional Xgboost model or conventional classifiers such as KNN, decision trees, random forest.

C. A hybrid model fro predicting mobile price range

In [3], various classification methods were employed to predict mobile price ranges. The initial approaches involved the utilization of Decision Tree and Random Forest machine learning algorithms, yielding accuracies of 83% and 84%, respectively. To enhance the accuracy of the Decision Tree model, parameter pruning was implemented, resulting in a subsequent accuracy boost to 90% when Random Forest was applied. Additionally, a hybrid ensemble method was introduced, incorporating five distinct heterogeneous weak learners, a strategy with a track record of demonstrating superior performance in machine learning. This approach had not been previously explored in the context of this dataset. Furthermore, a comprehensive performance evaluation was conducted for Decision Tree, Random Forest, and the hybrid model, considering metrics such as Precision and Recall.

D. Supervised learning algorithms for mobile price classification

As we can see from [4], This study employs a range of classification methods, including stochastic gradient descent, random forest, KNN, support vector machine (SVC), naive Bayes, artificial neural network (ANN), decision trees, and logistic regression. The aim is to identify and eliminate unnecessary features while minimizing computational complexities. Multiple classifiers are utilized to achieve superior accuracy, and the results are compared based on precision. Finally, a conclusion is drawn, identifying the optimal characteristic classifier for a given dataset. In this case, SVC, ANN, and logistic regression demonstrate the best accuracy and results across all four classes.

E. Feature Engineering and Model Optimization Based Classification Method for Network Intrusion Detection

Paper [5] deals with classification and preprocessing. Given the increasing prevalence of cyber threats in the expanding landscape of the Internet, traditional machine learning methods face challenges in effectively detecting and classifying diverse cyber-attacks.[5] proposes a novel intrusion detection classification approach that combines advanced feature engineering and model optimization techniques. By utilizing mutual information maximum

correlation minimum redundancy (mRMR) feature selection and synthetic minority class oversampling technique (SMOTE), the method processes network data, addressing issues like feature redundancy and imbalanced class distribution. The study also employs the Optuna method to fine-tune hyperparameters for the Catboost classifier, enhancing overall model performance. Experiments using NSL_KDD, UNSW-NB15, and CICIDS-2017 datasets demonstrate the superiority of the proposed method in accuracy, recall, precision, and F-value over traditional approaches, showcasing its potential for effective network intrusion detection.

F. Classification Algorithm and Evaluation metrics of Machine Learning

As seen in [6] classification is an important part of machine learning. Machine learning relies significantly on classification, with notable advancements in algorithms and their applications. This article presents an overview of five fundamental single classifier models for addressing classification challenges. It begins by elucidating the core concepts of each model. Subsequently, it compares the strengths and weaknesses of each model, along with their applicable scenarios. The article then delves into the evaluation methods for classifiers and the performance metrics associated with them. Lastly, it analyzes the current status of development and identifies bottlenecks in the basic classification models.

G. House price prediction using ML

[7] employs similar type of classification techniques. This study addresses the challenge of predicting changes in house prices by framing it as a classification problem and leveraging machine learning techniques. Various methods, including variance influence factor, Information value, principal component analysis, and data transformation techniques such as outlier and missing value treatment, as well as box-cox transformation, are employed for feature selection. The performance of the machine learning models is assessed using four key parameters: accuracy, precision, specificity, and sensitivity. The study categorizes house price changes into two classes, represented by discrete values 0 and 1. A class value of 0 indicates a decrease in house price, 1 indicates an increase in house price.

H. Feature selection based on a mutual information measure for image classification

As seen in [8] efficient classification of hyperspectral images is a significant challenge due to their extensive information content. This paper addresses this challenge by analyzing the information in each spectral band and introducing an enhanced feature selection technique. The proposed method aims to minimize dependent information while maximizing relevancy, leveraging normalized mutual information (NMI). The paper includes experimental results comparing the classification accuracy of the proposed technique with other recent hyperspectral feature selection methods, using real hyperspectral images.

I. Mobile model price prediction using K means

In [9] K means was used to predict the price segment a mobile model belongs to. A real dataset was collected from the website www.GSMArena.com. Various feature selection algorithms were employed to identify and eliminate less important and redundant features with minimal computational complexity. Different classifiers were then applied with the aim of achieving the highest possible accuracy. The results were compared based on the maximum accuracy achieved and the minimum number of features selected. Conclusions were drawn regarding the optimal feature selection algorithm and classifier for the given dataset.

J. House resale price prediction using classification algorithms.

This [10] paper explores the prediction of house resale prices through the utilization of various classification algorithms, including Logistic Regression, Decision Tree, Naive Bayes, and Random Forest. The AdaBoost algorithm is employed to enhance weak learners to strong learners. The analysis considers multiple factors influencing house resale prices, such as physical attributes, location, and economic factors prevailing at the time. Accuracy serves as the performance metric for different datasets, and the algorithms are applied and compared to identify the most suitable method that sellers can refer to when determining resale prices.

III. Research Gaps And Objectives

A. Research gaps

- Preprocessing Efficiency: Current preprocessing techniques may lack optimization for large datasets and intricate feature spaces, leading to potential computational bottlenecks. Gaps in preprocessing efficiency must be investigated to ensure swift and effective data preparation.
- Feature selection: Assess the robustness of existing feature selection methods in capturing relevant features while minimizing redundancy. Identify gaps in these techniques that might compromise the model's ability to discriminate essential attributes for accurate price segment classification.
- Dynamic pricing factors: The existing models may not fully account for dynamic factors affeting mobile device prices, such as rapid technological advancements, changing market trends, and economic fluctuations. Addressing these are crucial for enhancing the model's accuracy over time.

B. Objectives

- Optimized preprocessing techniques: To develop preprocessing techniques which optimize efficiency without reducing data integrity. Also aims to enhance the speed and effectiveness of preprocessing steps for improved model performance.
- Improved feature selection methods: Our model aims to investigate and implement advanced feature selection methods that establish balance between information retrieval and dimensionality reduction. It also addresses gaps identified in the robustness of current feature selection techniques.
- Adaptability: This model aims to explore techniques to enhance the classifier's adaptability across different datasets and scenarios and devise strategies to improve

the model's generalization capabilities, making it more versatile and applicable to a broader range of mobile device markets.

- Provide accurate classification results: This model aims to accurately classify a mobile device based on its specifications and visualize the results obtained.

IV. Methodology

- Data collection:
  Dataset [11] has been used for this project. The dataset consists of 20 distinct features of 1000 different mobile device models. The dataset consists of features like RAM, Battery power, pixel width and height, processor clock speed and a lot more.The devices are classified into 4 segments, segment 0 denotes the lowest price whereas segment 3 denotes the highest price.
- Data pre-processing:
  The dataset contains no missing data or character data, but the values of features are spread across a wide range. For instance, the value of the 'battery' feature ranges between 501 to 1998 whereas the 'blue' feature ranges between 0 and 1. This varying range can cause the training process to be inefficient.
  We use minmax normalization to tackle this problem, in this case, without normalization, the feature with the larger range of values will have significantly more weight in the training than the feature with low range, so min-max normalization is introduced to map all the data to the interval [0, 1]. This speeds up the convergence of the model and increases the accuracy of the results. Its formula is given by:

$$X'=(X-Min)/(Max-Min) \quad (1)$$

  Here min is the minimum value of the feature and max is the maximum value of the feature, X is the value of the feature chosen currently.
- Feature engineering:
  The normalized dataset still consists of 20 features, passing all features to a classifier will result in high training time, overfitting. In order to prevent this, only features that have a significant role in predicting class must be selected for training the model.
  Feature importance was calculated using an algorithm called Mutual Information-Based Maximum Feature Minimum Redundancy (mRMR) Feature Selection. The mRMR algorithm, founded on mutual information, captures the complex non-linear relationships among the features. By utilizing information theory principles like information entropy, information gain, and mutual information, the MRMR algorithm gauges the correlation between features and target variables. The algorithm quantifies classification performance by evaluating the magnitude of the mRMR value, which signifies the strength of the relationship between features and class labels.
  The mRMR calculation is as follows:

$$I(X,Y)=\sum_{x,y} p(x,y) log p((x,y)/(p(x)p(y))) \quad (2)$$

  Here I(X,Y) is the amount of mutual information shared between 2 statistically independent features X,Y .The final mRMR value is computed by the following equation ;

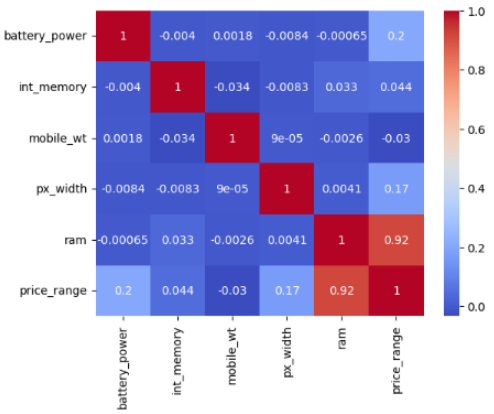
$$maxJ(D,R), J=D-R \quad (3)$$

  Here D denotes the maximum level of correlation between a feature and target variable and R denotes the minimum level of redundancy between features. Finally, the features which have maximum correlation and minimum redundancy are selected. Their equations are as follows:

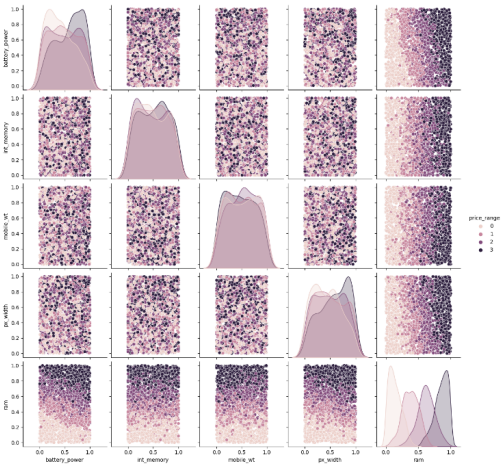
$$D=(1/|S|)\sum f_{i\in S} I(fi,c) \quad (4)$$

$$R=(1/|S|^2)\sum_{fifj\in S} I(fi,fj) \quad (5)$$

  Here S denotes feature space, f denotes feature and c denotes the class.
  The features chosen are namely ram, battery_power, px_width, mobile_wt, int_memory
- Visualization:
  The correlation matrix of the chosen features is visualized.

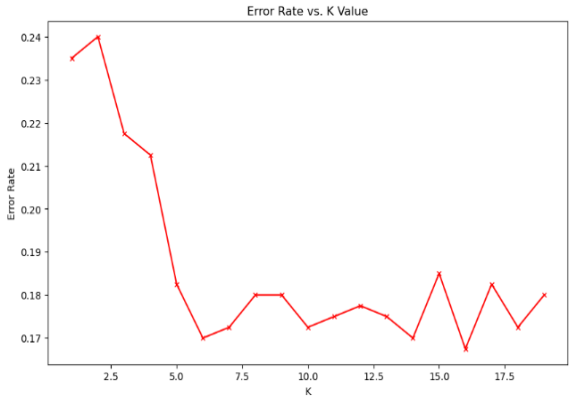


  Further, a pairplot is plotted to visualize the relation between the selected features and target variable.



- Choosing classifier:
  In order to choose the best model we have compared namely KNN, Logistic regression classifier, decision tree classifier, random forest classifier and naïve bayes classifier.

  KNN: KNN is a non-parametric, instance-based learning algorithm that classifies new data points based on their similarity to existing data points. It does not assume any underlying relationship between the data points, and it is therefore well-suited for data that is difficult to model with traditional parametric methods.
  The KNN algorithm works by identifying the k nearest neighbors of a new data point in the training set. The class of the new data point is then determined by the majority class among its k nearest neighbors. The value of k is a hyperparameter that can be tuned to optimize the algorithm's performance.
  In order to find optimal K we have used the elbow method, the results from the same are as follows:



  Minkowski distance metric was used to calculate the distances.

  Multiclass logistic regression: Logistic Regression is a classification algorithm that models the probability that a given instance belongs to a particular class.
  The logistic regression algorithm works by iteratively updating the weights of the linear function until it minimizes a loss function. The loss function is typically the cross-entropy loss, which measures the discrepancy between the predicted probabilities and the actual labels of the training data. The logistic function (sigmoid function) is commonly used. The logistic function is defined as:

$$P = 1/(1+e^{-(b_0 + b_1x_1+b_2x_2+....)}) \quad (6)$$

  Decision tree classifier: A decision tree is a tree-like structure that represents a set of decisions and their possible consequences. It is a classification algorithm that can be used to predict the class of a new data point based on its features.

  The decision tree algorithm works by recursively partitioning the training data into smaller subsets. At each partition, the algorithm selects a feature and a threshold value to split the data into two subsets. The process continues until each subset reaches a stopping criterion, such as a maximum depth or a minimum number of data points.
  The splitting of nodes is based on criteria like Gini impurity or entropy. The decision at each node is determined by finding the feature and threshold that best separates the data.
  To select the best feature the following entropy function is used:

$$\text{Entropy} = \sum -p_i * \log_2(p_i) \quad (7)$$

  Here pi is the proportion of class i in the dataset. From the value of entropy obtained we can calculate the information gain value of an attribute.

$$\text{Gain} = \text{Entropy} - \sum(s_v/s) * \text{entropy}(s_v) \quad (8)$$

  Attribute with the highest information gain value is chosen to carry out splitting.

- Random forest classifier: A random forest is a collection of decision trees. It is a classification algorithm that can be used to improve the accuracy of decision trees by reducing their variance.
  The random forest algorithm works by training many decision trees on different subsets of the training data. Each tree is trained on a random subset of the data, and each feature is considered only a subset of the total features. This helps to reduce the correlation between the trees and make the forest less prone to overfitting.

- Naïve bayes classifier: Naïve Bayes is a probabilistic classifier based on Bayes' theorem with the "naive" assumption that features are conditionally independent given the class. The formula for the posterior probability in the context of classification is given by Bayes' theorem.

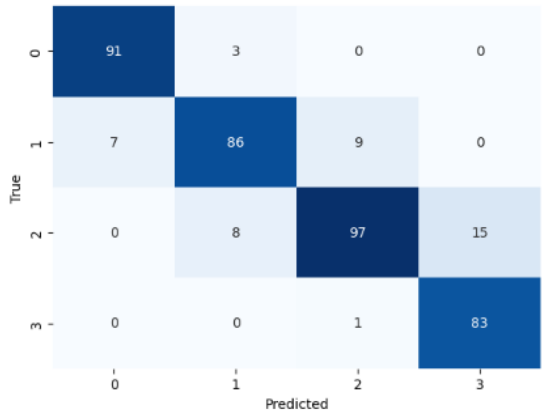
Post this the results obtained from all the models are compared and the best model is chosen.

V. Analysis of Results:

The scores from the above models are as shown below:

Model	Score
KNN	0.8325
Logistic regression	0.8925
Decision Tree	0.815
Random forest	0.8625
Naïve Bayes	0.75

It can be seen that Logistic regression works the best. Hence classification is carried out using logistic regression classifier. The visualization of the confusion matrix after logistic regression is as follows:



Various performance parameters like recall, F1 score, precision and support can be calculated from the confusion matrix. The results are as seen below:

Class	Precision	Recall	F1 score	Support
0	0.93	0.97	0.95	94
1	0.89	0.84	0.86	102
2	0.91	0.81	0.85	120
3	0.85	0.99	0.91	84

Overall accuracy is 89% using logistic regression.

VI. Conclusion and Future Work

The current model can efficiently classify mobile phones into price segments with an accuracy of 89% by using logistic regression, however the model cannot handle certain scenarios like :

- Dynamic Price changes: The current model does not take into account the change of trends in market, technological advancements and other economic fluctuations.
- Varying Interpretability : As the model complexity increases, there may be a gap in ensuring interpretability and explainability, which are crucial for user trust and understanding how the model arrives at its predictions.

Further improvements can be made to consider the dynamics of the market and customer choices as well to further improve the model.

References

[1] M. Çetın and Y. Koç, "Mobile Phone Price Class Prediction Using Different Classification Algorithms with Feature Selection and Parameter Optimization," 2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, 2021, pp. 483-487, doi: 10.1109/ISMSIT52890.2021.9604550.

[2] Y. Zhang, Q. Ding and C. Liu, "An Enhanced XGBoost Algorithm for Mobile Price Classification," 2023 IEEE 6th International Conference on Big Data and Artificial Intelligence (BDAI), Jiaxing, China, 2023, pp. 154-159, doi: 10.1109/BDAI59165.2023.10256847.

[3] A. H. Sakib, A. K. Shakir, S. Sutradhar, MD. A. Saleh, W. Akram and K. B. MD. B. Biplop, "A hybrid model for predicting Mobile Price Range using machine learning techniques", 2022 The 8th International Conference on Computing and Data Engineering, pp. 86-91, Jan. 2022. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[4] Dutta, A., Mallick, P.K., Mohanty, N., Srichandan, S. (2022). Supervised Learning Algorithms for Mobile Price Classification. In: Mallick, P.K., Bhoi, A.K., Barsocchi, P., de Albuquerque, V.H.C. (eds) Cognitive Informatics and Soft Computing.

[5] Zhang, Y.; Wang, Z. Feature Engineering and Model Optimization Based Classification Method for Network Intrusion Detection. Appl. Sci. 2023, 13, 9363. https://doi.org/10.3390/app13169363

[6] Z. Wu, J. Zhang and S. Hu, "Review on Classification Algorithm and Evaluation System of Machine Learning," 2020 13th International Conference on Intelligent Computation Technology and Automation (ICICTA), Xi'an, China, 2020, pp. 214-218, doi: 10.1109/ICICTA51737.2020.00052.

[7] D. Banerjee and S. Dutta, "Predicting the housing price direction using machine learning techniques," 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), Chennai, India, 2017, pp. 2998-3000, doi: 10.1109/ICPCSI.2017.8392275.

[8] M. A. Hossain, X. Jia and M. Pickering, "Improved feature selection based on a mutual information measure for hyperspectral image classification," 2012 IEEE International Geoscience and Remote Sensing Symposium, Munich, Germany, 2012, pp. 3058-3061, doi: 10.1109/IGARSS.2012.6350780.

[9] "Prediction of Mobile Model Price using Machine Learning Techniques", IJEAT, vol. 11, no. 1, pp. 273-275, Oct. 2021.

[10] P. Durganjali and M. V. Pujitha, "House Resale Price Prediction Using Classification Algorithms," 2019 International Conference on Smart Structures and Systems (ICSSS), Chennai, India, 2019, pp. 1-4, doi: 10.1109/ICSSS.2019.8882842.

[11] Mobile Price Prediction | Kaggle