Intel Unnati Industrial Training 2023

Smart Mobile Phone Price Prediction using ML

Team name: GO SOLO

Team Members

The project was carried out by the following member:

- Shaun Varghese (shaunvarghese@karunya.edu.in)

Mentors

I would like to extend my sincere gratitude to my mentors for providing constant support and guidance through live sessions and interactions:

- Dr. R. Chitra (Academic Mentor) - chirar@karunya.edu
- Mr. Srivatsa Sinha (External Mentor) - srivatsasinha@theprograms.in

Abstract

The use of smartphones has grown significantly during the last few decades. A smartphone is useful for everything from making grocery purchases to reserving airline tickets. Given the wide variation of smartphone pricing, it is critical to classify them and research their costs. This report makes an effort to research the various elements that influence smartphone prices and, using Dimensionality Reduction techniques, pick the best of their combinations.

1. Literature Survey and Related Work

Due to mobile phones' important role in communication, social networking, work, making payments, and entertainment during the past ten years, the market for them has grown rapidly. More and more businesses can now be conducted using mobile devices. The mobile phone penetration rate has increased quickly; in 2016, it was less than 50%, but by 2020, it had increased to 78.05%. [1]

To effectively manage high dimensional data, data reduction is required. Two well-liked methods of data reduction are feature selection and dimensionality reduction. The data size is decreased through feature selection by deleting unimportant features. Principal component analysis (PCA) is a method that is frequently used in cancer research to reduce dimensionality. PCA was used by authors in [2], [3] to minimize data size and produce significant analytical findings.

2. Dataset Selection and Exploratory Data Analysis

A mobile phone's brand, specifications, and features all have an impact on its cost. Pricing the newly introduced mobile phone reasonably in comparison to the competition can aid in its successful launch. The apt dataset for the above mentioned reasons can be found at [4].

Pre-processing is typically a procedure that comes after gathering data from the actual world in order to prepare it for data analysis. Pre-processing correctly could reduce the impact of data variability on the creation of parsimonious models.

A thorough data inspection reveals no noisy or missing attribute values in the provided dataset. The dataset is completely balanced, which will improve the accuracy of the model.

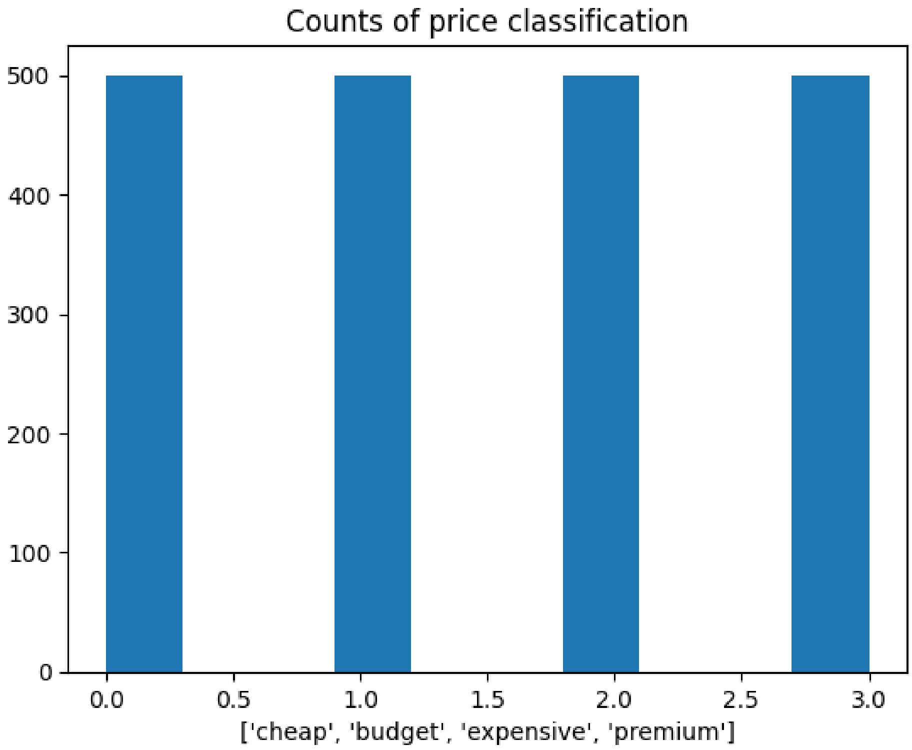

Fig 2.1: Instances for each price category

Each class has exactly 500 instances making it a perfectly balanced dataset. All the labels in the dataset are numerically encoded and hence it is ready for model building.

3. Method

3.1 Dimensionality Reduction

Dimensionality reduction is a method for reducing the number of features in a dataset while preserving as much crucial data as feasible. In other words, it is a method for converting high-dimensional data into a lower-dimensional space while keeping the original data's essential characteristics. High-dimensional data in machine learning is defined as data having a lot of characteristics or variables. A typical issue in machine learning is the "curse of dimensionality," where the model's efficiency degrades as the number of features rises. [5].

The report explores two popular dimensionality reduction, Principal Component Analysis and Linear Discriminant Analysis. The dataset is first built on a raw model, where there is no dimensionality reduction applied. Then the following dimensionality reduction techniques are applied.

3.2 Principal Component Analysis

An approach for an unsupervised learning algorithm called Principal Component Analysis (PCA) is used to look at how a group of variables are related to one another. Regression that chooses a line of best fit is referred to as a generic factor analysis. Without any prior knowledge of the target variables, the primary objective of PCA is to decrease the dimensionality of a dataset while maintaining the most significant patterns or correlations between the variables. This is accomplished by identifying a new set of variables that is both less comprehensive than the original set and more effective for the regression and categorization of data.

3.3 Linear Discriminant Analysis

A supervised learning approach called Linear Discriminant Analysis (LDA) is utilized in machine learning for classification applications. It is a method for determining the optimum linear combination of features for classifying a dataset's classes. In order for LDA to function, the data are projected onto a lower-dimensional space with a maximum degree of class separation. Finding a group of linear discriminants that maximize the proportion of between-class variation to within-class variance is how it accomplishes this. In other words, it determines the directions in the feature space that effectively distinguish the various data classes.

4. Model and Results

The dataset was trained initially without any dimensionality reduction on a raw Logistic Regression model. As the dataset was balanced and clean, an accuracy of 68.4%. All the 20 features of the dataset were used to train the model.

Then a model was trained using the Principal Component Analysis technique, in which 14 of the 20 features were used. This model yielded an accuracy of 80.2 %, which was around 12% increase from the raw model.

To further increase the accuracy of the model, Linear Discriminant Analysis was applied to the training dataset which uses the best combination of the 20 features. This model yielded a whooping accuracy of 93.2%, which is an increase of 13% and 25% increase from the previous two models respectively.

The following table summarizes the accuracy of the three models.

Table 4.1: Accuracies of the models

Model	Accuracy
Raw Model	68.4%
Model with PCA	80.2%
Model with LDA	90.2%

Hence, the model with the highest accuracy, i.e the model with Linear Discriminant Analysis was used to carry out the tests on the test dataset from Kaggle. [6]

Conclusion

So, the dimensionality reduction can greatly affect the predictions and accuracy of an ML model. The train dataset [4] was trained with different models and the best one (LDA) was picked to make predictions on the test dataset [6].

GitHub Repo link:
https://github.com/iamshaun08/Intel-Unnati-Industrial-Training-Summer-2023

Colab Notebook link:
https://colab.research.google.com/drive/1CN1AW2dC0OzZWLXf_zPOO90GMroc-8Ux?usp=sharing

References

[1] https://www.statista.com/statistics/263437/global-smartphone-sales-to-end-users-since 2007/

[2] Chiu H.-J., Li T.-H.S., Kuo P.-H., Breast cancer–detection system using PCA, multilayer perceptron, transfer learning, and support vector machine

[3] Adiwijaya W.U., Lisnawati E., Aditsania A., Kusumo D.S., et al. Dimensionality reduction using principal component analysis for cancer detection based on microarray data classification

[4]https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification?resource=download&select=train.csv

[5] geeksforgeeks.org/dimensionality-reduction/

[6]https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification?resource=download&select=test.csv