Business Intelligence

Assignment 2 - Interim Submission

Bakir Bajrovic - Person A - 12347510
TU WIEN
Vienna, Austria

Emirhan Kurtulus - Person B - 12243493
TU WIEN
Vienna, Austria

Abstract

In this assignment, we represent an analysis for a business intelligence task by using dataset from Kaggle[3], which contains data mobile phone features data by defining the price range according to these specifications. In this project we followed the CRISP-DM methodology to develop a predictive model. From the business objectives focus on increasing the sales, maximizing the market share, and identifying the important features for defining the price range for it. As the data mining goal, as we mentioned before we try to define the price range based on the features such as RAM, and battery power of a phone. Moreover, the quality of the data is checked and taken some preprocessing steps to prepare the data for modeling part. Potential risks related to AI explainability and data bias were valued. To have an insights on the data some visualization techniques were used such as line charts, scatter plots. With this interim report we illustrated that we have found the initial findings and the steps for the data preparation while preparing the stages for the further modeling, development, and evaluation.

Keywords

Business Understanding, Classification Task, Data Mining, Machine Learning, Crisp-DM

Bakir Bajrovic - Person A - 12347510 and Emirhan Kurtulus - Person B - 12243493. 2024. Business Intelligence: Assignment 2 - Interim Submission.

1 Business Understanding

1.1 Define and Describe the Data Source and Scenario

The data source for this assignment is the Mobile Price Classification dataset from Kaggle, that Bob and his company collected (see dataset description[3]), which contains information about 2000 previous mobile phone sales collected from the market (regional or global, not specified), together with the phones' characteristics/specifications, and the price range which their selling price belongs to.

In a real-world scenario, this dataset is meant to help mobile phone manufacturing companies to identify key features that determine a mobile phone's price range, and also to classify which price range a phone belongs to based on its specifications. This is important for companies competing on the market with other, possibly more successful companies. It might assist them with getting a better understanding of the current state of the market and identifying which features of a phone are most important and in demand, while also being able to decide on a fair price range to get an optimal value for money ratio that would be appealing to the customers.

Assignment 2, November 2024, Vienna, Vienna, Austria

1.2 Business Objectives

The primary objective is for Bob's company, which is in our case our customer, to catch up with the big companies and give them competition, by improving his company's performance with some better data analysis. To "catch up with the big companies and improve performance" would mean to **increase sales**, **maximize market share** and **make more profit**. The next few objectives are there in order to accomplish the first one. The second objective is to build a good model that is able to **predict a fair price range** for a mobile phone with certain characteristics. The third, also very important objective, is to **identify the key phone features** that have the biggest impact on the price range, and basically see how the phones in different price ranges differ, which specifications and types of phone configuration are associated with each price range. These goals, if met, should increase user satisfaction, leading to more people buying mobile phones from this company. As for the AI regulations and risk classification, in our case there is minimal or no risk, since we are only dealing with data regarding mobile phones, which does not directly impact any rights or safety of an individual.

1.3 Business Success Criteria

To be able to see if the business objectives are met in the end, some business success criteria is defined:

- Increase sales by 10% within the first 6 months
- Increase market share by 5% within the first 6 months
- Develop at least 1 mobile phone for each price range that best represents/fits that category, taking into account the best configurations found in the process

This criteria is assessed by the business owner and company.

1.4 Data Mining Goals

The goal of our data mining task is to create a classification model which predicts the **price range** of mobile phones based on its technical qualifications, in four different categories from 0 to 3. In the task, we aimed at "predicting fair prices" by using purchasing features like RAM, camera quality, number of pixels for the screen (by width, and height), and battery power, predicting the price range of that phone. Moreover, we also aimed to determine which features have an impact to define the price. This may help us to determine the value of a phone.

1.5 Data Mining Success Criteria

For the scope of success criteria, it will be measured by achieving accuracy with the score that is at least 90%, with other additional success criteria such as precision and recall to reduce the misclassifications by evaluating the results.

Another important point to point out that model is its explainability, which will help people to understand the reasons behind the predictions. Thanks to the fact that we have four different classes in our target variable, it is convenient to apply classification models such as k-nearest neighbors (k-NN), or random forest. Furthermore, the analysis will identify at least one key feature for each price range. For example, all expensive phones might share a high-quality camera feature, highlighting essential attributes for each category.

1.6 AI Risk Aspects

Our dataset does not indeed present specific AI risk aspects, particularly taking into account its explainability and transparency requirements. According to the paper that we utilized, the European Union's AI Act[2], it illustrates that datasets contributing to high-risk AI must meet strict transparency, traceability, and data quality requirements. With this approach it is aimed to ensure the explainability of models trained on such data, which is critical for maintaining accountability and user trust by giving the accurate results for them. Moreover, challenges that we might come across around fairness and bias in dataset used for price prediction classification models. As mentioned in the paper, suggests that organization should consider the model's accuracy vs. explainability trade-offs.[2] We have not taken the modelling steps yet, and in our case we have just the features of a phone which is referring any personal data, even if we would have done this part that might lead to misclassification for the price range that can just cause to choose less important features for a phone or less profit for the seller.

2 Data Understanding

2.1 Attribute Types

The dataset has 21 columns, where 20 of them represent certain characteristics or properties of a mobile phone, whether it has them or not, or which are the exact values of those attributes. All attributes are in numerical format, either integer or floating point values. The final column is the target variable, price range, also of type integer. The columns of this dataset are:

- **Battery_power** - Total energy of a battery that can be stored after full charge, in mAh (*integer*)
- **Blue** - A categorical variable that is label encoded, shows whether a phone has Bluetooth or not (1 - yes, 0 - no, *integer*)
- **Clock_speed** - Speed at which the microprocessor executes instructions, displayed in decimal values (*float*)
- **Dual_sim** - Another categorical value that was already label encoded (1 - has dual SIM support, 0 - does not have, *integer*)
- **Fc** - Shows the number of megapixels that the front camera has (*integer*)
- **Four_g** - (1 - has 4G, 0 - does not have, *integer*)
- **Int_memory** - Internal memory of a phone in GB (*integer*)
- **M_dep** - Mobile depth in cm (*float*)
- **Mobile_wt** - Weight of a mobile phone, in grams (*integer*)
- **N_cores** - Number of cores of a processor inside the phone, has distinct numerical values (*integer*)
- **Pc** - The number of megapixels of a primary camera (*integer*)
- **Px_height** - Pixel resolution height (*integer*)
- **Px_width** - Pixel resolution width (*integer*)
- **Ram** - Amount of RAM memory in MB (*integer*)
- **Sc_h** - Screen height of a phone in cm (*integer*)
- **Sc_w** - Screen width, also in cm (*integer*)
- **Talk_time** - Longest time that a single battery charge can last while on a call (*integer*)
- **Three_g** - (1 - has 3G, 0 - does not have, *integer*)
- **Touch_screen** - (1 - has touch screen, 0 - does not have, *integer*)
- **Wifi** - (1 - has WiFi, 0 - does not have, *integer*)
- **Price_range** - This is the target variable, consisting of 4 different classes (0,1,2,3) representing 4 price ranges. The exact price ranges in terms of some monetary value are not given or unknown, but this helps to at least cluster some phones to see which ones should be given a similar price.

2.2 Statistical Properties

The dataset that will be used consists of 2000 rows/instances. This is marked as the train dataset, and while a test set is also provided on Kaggle, it will not be used since there are no available price ranges for the instances in the test set that we can use for comparison and evaluation of our results. This is why only the "train" set will be used, and later split into our own train and test subsets. The dataset has 21 columns including the target variable, as previously mentioned. Most of the columns of the dataset have different scales, which is to be expected (battery power has a maximum value of 1998 and minimum of 501, while clock speed has max 3 and min 0.5), since different properties of the phones use different measures and measurement units. For analyzing the statistical properties in depth, we plotted a box plot for each column/feature to see the distribution, median, min, max and the interquartile ranges, while also checking for potential outliers. The df.describe method was also used to get a good overview of the values of the dataset and see the exact values of all these statistical properties. This is all shown in the supplementary notebook file[1]. Another interesting aspect is the correlation between the columns, especially between the target variable and all the others from the Figure 1.

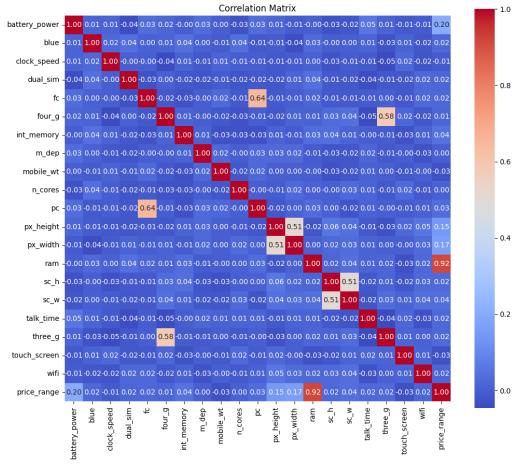

Figure 1: Correlation Matrix for the Features

2.3 Data Quality Aspects

The quality of the dataset is very good overall, since it seems like it was already prepared before by the owner. There are no missing values anywhere and all attributes are already converted into a numerical format which is important for the ML models. Most of the attributes are normally distributed, except for some which are slightly right skewed (like px_height and sc_w). This is something that might have to be examined further if the skewness of these attributes becomes an issue down the road. Also, the boxplots from before show just a few data points considered as attributes, but they do not seem to be extreme. The target variable is evenly distributed across all classes, where each class is represented 500 times (across 2000 rows), which is basically ideal for a classification task.

2.4 Visual Exploration

In this section, we tried different visualization approaches to gain more insight into our dataset and its attributes. First, we visualized several pairwise scatterplots to see the correlation between some attributes, as well as to try and identify some patterns or clusters and see if the classes are clearly divided or is there some overlapping.

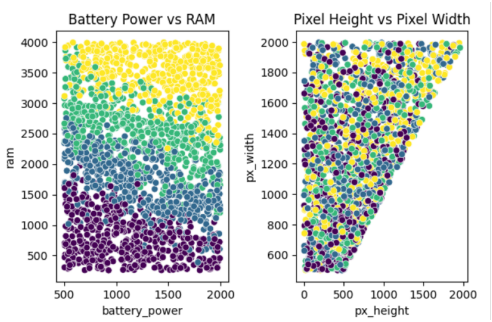

Figure 2: Pairwise Scatterplots

In the figure 2 above only two scatterplots are shown, since the other two that were created were not too clear or informative. One of them showed the relationship between the front camera and primary camera, and there the classes were completely scattered, while many data points were on top of each other because they had same values, so it appeared as there are less data points than there actually are. The second plot that was omitted was the one that displayed 4G vs 3G, which had an even more extreme case of data points overlapping.

As for the other two plots that are visible on figure 2, the first scatterplot shows nicely how the classes (or price ranges) are distributed over RAM and battery power. The classes are clearly visible and we can notice **4 clusters** formed, one for each class, with the exception of few data points where some neighboring classes are overlapping, which is expected. This is a good indicator that as the RAM increases, the price range of the phones also tends to increase. The same is with battery power, but less impactful, since we can see that for every battery power value, there are some data points representing each class, but overall out of the phones with the highest battery power, most of them belong to price range 3. The second scatterplot does not show any pattern, which is also a useful observation, meaning that the pixel height and width cannot help that much with determining a phone's price range. Better camera specifications do not guarantee a higher price range for a phone, which was not our initial hypothesis, but an interesting finding nonetheless.

The second part of the visual exploration involved analyzing the distribution of the binary attributes per price range, to try to see if the distributions are different across the different price ranges. This is shown in the figure 3 below.

Figure 3: Distribution of Binary Attributes per Price Range

In figure 3, we also arrive to an interesting and important conclusion. Seeing that the distributions across different price ranges are almost the same for every attribute, where the ratio between 1 and 0 is very close to (or even exactly) 50-50 in all cases. This basically means that none of the binary attributes can really help us in determining a phone's price range, since none of them are unique characteristics of a certain price range.

The final plot that is part of the visual exploration of this dataset is a plot showing the attribute means per price range.

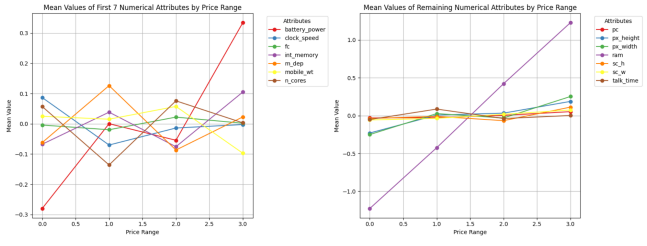

Figure 4: Attribute Means per Price Range

We used this plot to try and identify how the average values of the numerical attributes change over different price ranges, to try to identify some patterns, whether some attribute values increase or decrease, so that it is easier to detect which price range a phone could potentially belong to. Because there were 14 numerical attributes, we split them into two separate plots for better visibility and easier color coding, since it was too hard to find 14 distinct colors to show and still be able to see on one plot.

The data also needed to be scaled before creating this plot, so that it would be more zoomed in and each line can be more visible. This plot confirmed what was already seen in the correlation matrix and the pairwise scatterplot, and that is the fact that as the RAM increases, the price of the phone increases as well. Every next price range has a higher mean RAM than the previous one. A similar pattern can be noticed with the battery power, pixel height and pixel width. Another useful observation is that it is in most cases the hardest to distinguish between price ranges 1 and 2, because for many attributes the means are very close for these two classes. As for the lowest price range, it has the highest clock speed mean, so that is something that class can be distinguished by. The second price range has the highest average mobile depth, while the third has the highest average mobile width and front camera size. The priciest phones also seem to be the narrowest on average (lowest avg mobile width) which could also be a good indicator when identifying a phone with the highest price range. All of these insights can potentially help us to better distinguish which price range a certain phone belongs to, and to check whether our model makes a reasonable prediction later on.

We also tried using the parallel coordinates and PCA, but these types of visualizations did not provide any useful information worth mentioning.

2.5 Evaluation of potential presence of ethically sensitive data

Since the dataset we are dealing with consists of attributes and characteristics related to different mobile phones. There is no personal information or demographic data about any of the customers or employers, or even the names of the phones or the companies that make them. There should be no issues with privacy, social bias or any kind of discrimination. Another important thing is that the dataset is well balanced, where each price range contains an equal number of entries, which reduces the need for over- or under-sampling. The only potential problem could be that some groups of phones (specific combinations of attribute values, specific configuration) may not be represented enough.

2.6 Risk and types of bias

Even though the dataset does not contain any social data and a good structure in general, with most of its columns normally distributed, we can still point out some potential risks and types of bias. One could argue that the high correlations between some features (especially between features and target variable, like RAM) could introduce a bias that favors specific phone settings more and focuses too much on just a couple of attributes and certain attribute combinations, but it is not too likely that this will be a big issue. There is also potential for representation bias, because we do not know where the data was collected, is it related to a specific **region** or was it sampled from a more world-wide market. Also, how many different phone types were sampled, from which companies, how old is the data, this is all important information that is not shown in the dataset. It would be helpful if an expert could answer these questions by examining the data and potentially identifying some mobile phones based on the specifications, and if the data is consistent across different markets and "global" enough, plus if all the different phone types are represented enough.

2.7 Actions likely required in the preprocessing steps

One action that will most likely be done is applying one of the scaling methods, since it was already discussed that many attributes have very different scales, which could affect the performance of some classification algorithms such as K nearest neighbours, which rely on calculating distances between data points. There are very few potential outliers (in fc and px_height) that could be dealt with as well, but do not seem too extreme.

3 Data Preparation Report

3.1 Pre-processing Steps

First we check if there are any missing values, and we perform it, then we have seen that there are no missing values in the dataset. Then we check for the outliers in it, yet before that we check for the data types as we mentioned before all the data in the numerical format even though some of them are shown as binary. Next, we checked the numerical features to be sure which features to be considered, after that we illustrated the features to see the range of them, and we took the steps for visualizations to have a better understanding on the data. Moreover, before the scaling part we split the data as train and test sets, by excluding the target variable from the training set, we prepared the data for the further steps of Crisp-DM, and we checked for the features. As we see in the notebook[1] the range of features is different, to have more accurate results of the data we scaled by using the standard scaler from the sklearn library, because some of the features were between 500 and 2000, whereas there are some features in the range 0 and 20. By scaling these features we have a better understanding about the dataset, and more accurate values for the modeling part. Next, we did some explorations as we mentioned in the data exploration part.

3.2 Pre-processing Steps Not Applied

There were not any missing values, when we checked the missing values we saw that the data was collected in a good way, and distributed equally, but if we would have any missing value we needed to perform some preprocessing steps to handle them, one of the methods that is removing the missing values which leads information loss, or lower statistical power. Moreover, thanks to the dataset is equally distributed and there is no data out of the scope, and there is no outlier to disrupt the modeling part that is why we did not perform any method to eliminate the missing values in our dataset. For the data transformation part, there is no categorical value that data transformation should have been applied to, so we did not perform any encoding techniques such as One-hot-encoding, or Label encoding. Because all our features are numerical (binary, or in a range). Additionally, we have not done the log transformation, because we already used the standard scaler, and are dealing with a classification problem, some of the right-skewed attributes will not be a problem down the line and log transformation is therefore not necessary at this moment.

3.3 Potential for Derived Attributes

For attribute derivation we tried to use Pythagorean theorem to calculate the diagonal screen size from the attributes that define the screen which are px_height and px_width. But when we checked the impact of it to the price_range, there was no impact on it. That is why we decided not to use it as a derived attribute. Moreover, there might be other attribute derivations for further analysis, for example, the relationship between int_memory, and ram could be interesting to check, and also battery and screen size may be interesting to explore.

4 Additional External Sources

Attributes that reflect the purchasing power of the target market could provide valuable context for setting optimal price ranges. For example, pricing a luxury phone across different countries may not be feasible, as affordability varies significantly between high-income countries and those with developing economies which have a big amount of taxes on the phone such as Turkey. In summary, standardized prices across all regions may not align with local economic conditions and could impact the company's sales potential.

Acknowledgments

You can access the mobile price classification dataset on Kaggle from the link [3], and the repository, which is available publicly in Github [1].

References

[1] Emirhan Kurtulus Bakir Bajrovic. 2024. Business Intelligence Assignment 2. https://github.com/Emir515/Business-Intelligence-Assignment-2 Accessed: 2024-11-13.

[2] Luca Nannini, Agathe Balayn, and Adam Leon Smith. 2023. Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1198–1212. https://doi.org/10.1145/3593013.3594074

[3] Abhishek Official. 2023. Mobile Price Classification Dataset. https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification Accessed: 2023-11-13.