Feature Selection and Model Tuning for Mobile Features Data

Featurization and model tuning

Domain

Mobile Electronics

Context

Predicting the price of a mobile phone based on historical data as well as features can be a useful use case for mobile companies. Using the features that are part of a mobile phone, one can exploit the data that is already available and regress it to find the prices of future mobile prices. Also, it can be useful for determining whether a phone is rightly priced or not. However, it is important to understand which features impact the performance of our model and also what parameters of our model when tuned efficiently give a good overall performance for our predictions. Thus, understanding the features and model parameters is essential for the model building process for optimal performance.

Approach

To solve this problem data has been collected for information on mobile phones from various telecom companies.This data comprises technical specifications for different mobile phones along with its approximate price as the target variable.

Here, the data collected is from various telecom firms. Target column has different prices in euros for the mobile phones.

Attribute Information

- **brand**: Brand or company name of the phone manufacturer
- **model** :Model name of the phone
- **network_technology** Type network supported by model like GSM,LTE
- **2G_bands** :Supporting 2g bands
- **3G_bands** :Supporting 3g bands
- **4G_bands** :Supporting 4g bands
- **network_speed** :Speed of the supporting network
- **GPRS** :GPRS supported YES or NO
- **EDGE:** EDGE technology supported YES or NO
- **announced:** date of announcement
- **status** : Phone is available or not for purchase
- **dimentions:** dimensions of the phone height,weight, width
- **weight_g:** Phone weight in grams
- **weight_oz:** Phone weight in ounce
- **SIM:** Type of SIM supported Mini-SIM,Micro-Sim..
- **display_type:** Type of display LCD,LED
- **display_resolution:** Phone Resolution
- **display_size:** Phone display size
- **OS:** OS name
- **CPU:** Type of CPU
- **Chipset:** Chipset used in phone
- **GPU:** GPU used in phone
- **memory_card:** memory_Card size
- **internal_memory:** internal memory
- **RAM:** size of RAM
- **primary_camera:** Primary camera MP
- **secondary_camera:** secondary camera MP
- **loud_speaker:** loud speaker support or not
- **audio_jack:** Audio_jack Yes or NO
- **WLAN**: WIFI , Hotspot supported or not
- **bluetooth:** bluetooth supported or not

- **GPS:** Phone GPS
- **NFC:** Near Field communication
- **radio:** FM radio supported or not
- **USB:** USB Type of USB
- **sensors** :Type of Sensors used
- **battery:** Battery size
- **colors:** available colors
- **approx_price_EUR:** Price in euros
- **img_url:** image url

Steps

1. Once we have read the data into a dataframe, let's check the shape, information, datatypes and other attributes of the dataset.

2. Let us find the null values/ missing values in the dataset and get the total number of missing values per column

3. We can check the summary statistics of the dataset using the describe method in pandas.

4. Replace any special characters with NaNs

5. Let us group our dataset by brand using pandas groupby method

6. Impute missing data using already available data in the columns

7. Drop columns that do not have relevance with the target variable

8. Carry out Shapiro wilkinson test and comment your findings

9. One hot encode the categorical variables

10. Plot a histogram and correlation plot of all the features

11. Preparing the dataset by doing a train test split and scaling the features

12. Try out different models such as linear regression, random forest regressor and use recursive feature elimination in order to find the ranks of different features in our dataset. A rank of 1 would mean the feature is selected
13. Use ensembling methods such as Decision tree regressor, gradient boosting, random forest, adaboost and extra trees regressor to get a better understanding of the feature importances
14. Use GridSearchCV and RandomSearch CV to find optimal parameters in the RandomForestRegressor model that has been fitted onto this dataset
15. Visualize your models using scatter plots and residual plots. Infer your conclusion from the experiments carried out

Food for thought!

What more techniques can you try for finding the optimal hyper parameters?

Can you try and achieve a better performance than the one in the notebook?

Try out techniques such as Forward Selection and Backward Elimination for feature selection

Learning Outcomes

- Exploratory Data Analysis - dealing with mixed data, impute missing data
- Building ML models for regression
- Feature importance
- Hyperparameter Tuning