Mobile Phone Price Class Prediction Using Different Classification Algorithms with Feature Selection and Parameter Optimization

Mustafa ÇETİN
Istanbul Technical University
Information Institute
Istanbul, Turkey
cetinus@itu.edu.tr

Yunus KOÇ
Istanbul Technical University
Information Institute
Istanbul, Turkey
kocyun@itu.edu.tr

Abstract—Machine Learning (ML) algorithms are used in many fields such as finance, education, industry, medicine and e-commerce. ML algorithms show performance differences depending on the dataset and processing steps. Choosing the right algorithm, preprocessing and postprocessing methods has great importance to achieve good results. In this paper, Random Forest Classifier, Logistic Regression Classifier, Decision Tree Classifier, Linear Discriminant Analysis, K-Nearest Neighbor Classifier and SVC methods are compared to predict mobile phone price class. The “Mobile Price Classification” dataset which is taken from Kaggle.com is used to evaluate methods. Firstly, all values at the dataset are checked to verify that there are no missing values. After that, scaling is applied to dataset in order to obtain more relevant data for ML algorithms. Then, feature selection methods which reduce the computational cost by reducing the number of inputs are performed to get meaningful features. Finally, the parameters of classification algorithms are tuned to improve the system accuracy. According to obtained results, it is seen that ANOVA f-test feature selection method is more convenient for this dataset. It gives satisfying accuracy with a minimum number of features. It is also seen that the SVC classifier has the highest test accuracy compared to other models.

Keywords—Regression, Classification, Feature Selection, Parameter Optimization, ANOVA, Mutual Information, Random Forest Classifier, Logistic Regression Classifier, Decision Tree Classifier, Linear Discriminant Analysis, KNN Classifier, Support Vector Machine.

I. INTRODUCTION

Price is the most important side of marketing and shopping. Customers first would like to know whether the price of items is expensive or cheap [1]. Likewise, they want to know if the phone price is expensive compared to its features. For this, it is necessary to make price and performance comparisons between many phones and their features. But, making this search takes a very long time. However, Machine Learning (ML) algorithms can make a comprehensive comparison of phones and their features in a short time. They are used to find solution in many fields such as finance, education, industry, medicine and e-commerce [2,3]. ML algorithms, which can handle complex and large-scale data, assist many areas to make life easier. A well-designed ML algorithm gives satisfactory results even in complicated problems. There are a lot of studies in the literature that try to find the best models and parameters achieving high accuracy.

[4] proposes Naive Bayes Classifier and Decision Tree Classifier methods for mobile phone price class classification. Principal component analysis with forward and backward selection methods is applied to the dataset to remove insignificant features to minimize computational complexity. The study denotes that classification accuracy decreases by adding irrelevant features and removing important features. Decision Tree Classifier gives better results than Naive Bayes Classifier for this dataset. Artificial Neural Network is also a common used method to predict item price [5,6]. In [7], different classification algorithms are used to predict the house resale price. The paper makes comparison between Decision Tree Classifier, Naive Bayes Classifier, Ada Boost Classifier and Random Forest Classifier to find the most accurate method for determining house resale price. After the pre-processing steps, the dataset is evaluated by these four methods to get the best accuracy. Ada Boost classifier gives the best results for classification of this dataset. Another study for forecasting the housing price direction compares Artificial Neural Network, Support Vector Machine and Random Forest Classifier methods on the dataset taken from Kaggle.com. Variance influence factor, information value, principle component analysis and data transformation are utilized as feature selection techniques. Accuracy, precision, sensitivity and specificity metrics are calculated to measure performance of algorithms. Random Forest Classifier gives the highest accuracy, but it is affected over fitting problem. Support Vector Machine is seen as consistent and reliable compared to other methods [8]. In [9] and [10], Support Vector Machine method is used prediction of Bitcoin prices by taking the daily changes in the market of bitcoin. In [11], supervised learning techniques which are Decision Tree Classifier, Support Vector Machine Classifier and K-Nearest Neighbor methods are utilized to forecast gold price movement for determining right buying or selling time. Algorithms are learned by using only past gold prices as a feature. K-Nearest Neighbor algorithm gives sufficient accuracy while other methods do not. The used algorithms can give better results by using different financial parameters as features.

In this study, we compare Random Forest Classifier, Logistic Regression Classifier, Decision Tree Classifier, Linear Discriminant Analysis, K-Nearest Neighbor Classifier and SVC methods in order to predict mobile phone price class. The “Mobile Price Classification” dataset which is taken from Kaggle.com is used to evaluate performance of methods. The
dataset has 4 price classes and 20 features like ram, battery power, dual sim, memory, number of core, touch screen and screen size. ANOVA f-test measure and mutual information score are applied dataset for feature elimination. Moreover, hyperparameter optimization is carried out to improve system accuracy. At the end of the study, classifiers and feature selection methods are compared according to their classification performance. Hyperparameter is also examined for improving classification accuracy.

II. SYSTEM MODEL

Fig 1. shows main mobile phone price class prediction steps. At the first step, all values in the dataset are checked to verify that there are no missing values. It is seen that all values are valid. At the second step, standart scaling is applied to the dataset in order to obtain data more appropriate for classification algorithms. At the third step, Anova f-test measure and mutual information score feature selection methods are performed to select the best features. Feature elimination reduces computational cost by reducing the number of inputs. At the fourth step, the classification algorithms parameters are tuned to boost the performance of system. Finally, the dataset are classified into four class by using six different classification algorithms. Data scaling, feature selection and hyperparameter optimization steps are detailing described below.

[Diagram of the process]

A. Data Scaling

The purpose of data scaling is to transform the data to new form suitable for machine learning algorithms. Mining and modeling should be performed after scaling that is one of the preprocessing steps. Especially, algorithms using distance measures such as K-Nearest Neighbor Classifier need data scaling. Moreover, normalization of the data is an essential step for Artificial Neural Network algorithms [12]. The “Mobile Price Classification” dataset has mostly numerical features having different mean, standard deviation and range. Therefore, standard scaling is applied to numerical features.

\[
\bar{Y} = \frac{Y - \bar{Y}}{\sigma}.
\]

where \(Y\) is the original data, \(\bar{Y}\) is the mean of the feature, \(\sigma\) is the standard deviation of feature and \(\bar{Y}\) is the scaled data.

B. Feature Selection

Feature selection is the process of reducing the number of inputs while developing the predictive model. It reduces computational cost and sometimes improve model performance. There are different type of feature selection methods. Feature selection method should be selected according to data features and labels. “Mobile Price Classification” dataset has mostly numerical features and a categorical label. Therefore, the feature selection method should be appropriate for this input and output data types. ANOVA f-test measure and mutual information score can be used for the “Mobile Price Classification” dataset. Each method are described below.

1) ANOVA: Analysis of Variance (ANOVA) that is a statistical method determines the presence of variations between different population means. The method makes comparison among several types of variances [13]. The method uses the F-statistic indicating the feature significance. F-statistic is calculated as given below:

\[
F = \frac{\sum_{i=1}^{K} n_i (\bar{Y}_i - \bar{Y})^2}{K - 1} \cdot \frac{\sum_{i=1}^{K} \sum_{j=1}^{n_i} (\bar{Y}_{ij} - \bar{Y})^2}{N - K}.
\]

where \(\bar{Y}_i, n_i, \bar{Y}, \bar{Y}_{ij}, N\) and \(K\) represent the sample mean in the \(i^{th}\) feature, sample number of \(i^{th}\) feature, the average of the samples, \(j^{th}\) sample of \(i^{th}\) feature, feature number and the number of groups respectively.

2) Mutual Information: The mutual information is the measure of correlation between two variables and defined as [14]:

\[
I(X,Y) = \sum_{y \notin Y} \sum_{x \in X} p(x,y) \log \left( \frac{p(x,y)}{p(x)p(y)} \right)
\]

where \(X\) and \(Y\) are two random variables, \(p(x,y)\) is joint probability distribution, \(p(x)\) and \(p(y)\) are the marginal probability of \(x\) and \(y\).

ANOVA and Mutual Information feature selection methods are applied to the dataset and features are sorted from the highest score to lowest. Then, accuracy of each method is calculated by using the sorted features. Results given Table 1 and Table 2 are obtained for the number of selected best features being in second column. Features selected by ANOVA and Mutual Information have nearly same maximum accuracy for all algorithms, while ANOVA has lower number of features. Therefore, ANOVA is preferred as feature selection technique.
### TABLE I. PERFORMANCE OF ANOVA

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Number of Selected Best Features</th>
<th>Train Accuracy</th>
<th>Test Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest Classifier</td>
<td>4</td>
<td>100.0</td>
<td>91.14</td>
</tr>
<tr>
<td>Logistic Regression Classifier</td>
<td>7</td>
<td>84.48</td>
<td>83.75</td>
</tr>
<tr>
<td>Decision Tree Classifier</td>
<td>4</td>
<td>100.0</td>
<td>86.19</td>
</tr>
<tr>
<td>Linear Discriminant Analysis</td>
<td>5</td>
<td>95.27</td>
<td>94.94</td>
</tr>
<tr>
<td>KNN Classifier</td>
<td>4</td>
<td>93.47</td>
<td>89.24</td>
</tr>
<tr>
<td>SVC Classifier</td>
<td>4</td>
<td>96.03</td>
<td>93.95</td>
</tr>
</tbody>
</table>

### TABLE II. PERFORMANCE OF MUTUAL INFORMATION

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Number of Selected Best Features</th>
<th>Train Accuracy</th>
<th>Test Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest Classifier</td>
<td>5</td>
<td>100.0</td>
<td>91.60</td>
</tr>
<tr>
<td>Logistic Regression Classifier</td>
<td>9</td>
<td>84.48</td>
<td>83.94</td>
</tr>
<tr>
<td>Decision Tree Classifier</td>
<td>4</td>
<td>100.0</td>
<td>86.05</td>
</tr>
<tr>
<td>Linear Discriminant Analysis</td>
<td>8</td>
<td>95.22</td>
<td>95.00</td>
</tr>
<tr>
<td>KNN Classifier</td>
<td>4</td>
<td>93.41</td>
<td>89.34</td>
</tr>
<tr>
<td>SVC Classifier</td>
<td>4</td>
<td>95.89</td>
<td>94.1</td>
</tr>
</tbody>
</table>

### C. Hyperparameter Optimization

Hyperparameter optimization is used to control the learning process of algorithm. Hyperparameter optimization is an operation of choosing a set of optimum parameters for machine learning algorithm. The right method and hyperparameter selection has a vital role on the performance of classifier. Tuned hyperparameters yields an optimum model having higher accuracy. Manuel, grid or random search are used to tune hyperparameters [15]. In this method, grid search is used for hyperparameter optimization. Hyperparameters and search values are as below:

- **Random Forest Classifier**:
  - Number of estimators = [10, 50, 100, 250, 500]
  - Max depth = [50, 150, 250]
  - Min samples split = [2, 3]
  - Min samples leaf = [1, 2, 3]

- **Logistic Regression Classifier**:
  - C = [10-3, 10-2, 10-1, 0, 10, 100, 1000]
  - Penalty = [1, 10, 20, 30, 40, 50]

- **Decision Tree Classifier**:
  - Max depth = [50, 150, 250]
  - Min samples leaf = [1, 2, 3]
  - Min samples split = [2, 3]
  - Criterion = [gini, entropy]

- **Linear Discriminant Analysis**:
  - Solver = [svd, lsqr, eigen]

- **KNN Classifier**:
  - Number of neighbors = [2, 3, 4, 5, 6, 7, 8, 9, 10, 11]
  - Weights = [uniform, distance]
  - Algorithm type = [auto, ball_tree, kd_tree, brute]
  - Leaf size: = [1, 10, 20, 30, 40, 50]

- **SVC Classifier**:
  - C = [1, 10, 100, 1000]
  - Decision function = [ovo, ovr]
  - Gamma = [scale, auto]
  - Kernel function = [linear, rbf]

After grid search, the hyperparameters are obtained as below:

- **Random Forest Classifier**:
  - Number of estimators  =  500
  - Max depth = 150
  - Min samples split = 2
  - Min samples leaf = 1

- **Logistic Regression Classifier**:
  - C = 10
  - Penalty = 11

- **Decision Tree Classifier**:
  - Max depth = 250
  - Min samples leaf = 3
  - Min samples split = 3
  - Criterion = entropy

- **Linear Discriminant Analysis**:
  - Solver = svd

- **KNN Classifier**:
  - Number of neighbors = 10
  - Weights = distance
  - Algorithm type = auto
  - Leaf size: = 1

- **SVC Classifier**:
  - C = 10
  - Decision function = ovo
  - Gamma = scale
  - Kernel function = linear

Classification accuracies with tuned hyperparameters are compared accuracies before optimization. Results are shown in Table 3. Algorithms with optimized parameters give higher accuracy as seen in the table. Especially, accuracies of Logistic Regression and SVC Classifiers increase by a significant amount.
### TABLE III. ACCURACY AFTER HYPERPARAMETER OPTIMIZATION

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Before Hyperparameter Optimization</th>
<th>After Hyperparameter Optimization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train Accuracy</td>
<td>Test Accuracy</td>
</tr>
<tr>
<td>Random Forest Classifier</td>
<td>100.0</td>
<td>91.14</td>
</tr>
<tr>
<td>Logistic Regression Classifier</td>
<td>84.48</td>
<td>83.75</td>
</tr>
<tr>
<td>Decision Tree Classifier</td>
<td>100.0</td>
<td>86.19</td>
</tr>
<tr>
<td>Linear Discriminant Analysis</td>
<td>95.27</td>
<td>94.94</td>
</tr>
<tr>
<td>KNN Classifier</td>
<td>93.47</td>
<td>89.24</td>
</tr>
<tr>
<td>SVC Classifier</td>
<td>96.03</td>
<td>93.95</td>
</tr>
</tbody>
</table>

### D. Underfitting and Overfitting Analysis

Machine learning algorithms sometimes suffer from learning problems called underfitting and overfitting. Few number of features in training or simple models can lead to underfitting problem. Too many features in training or complex models can cause overfitting problem. It can be said that algorithms having much more training accuracy than test accuracy are affected by the overfitting problem [16].

Learning curves for each model are plotted to determine overfitting/underfitting case. Logistic Regression Classifier, Linear Discriminant Analysis and SVC models give good accuracies for both training and testing when the variance is not high. Moreover, increasing number of training data decrease the training accuracy while increasing test accuracy. So, there is no overfit/underfit case in these models. Random Forest Classifier, Decision Tree Classifier and KNN Classifier have high training accuracies while having low test accuracies relatively. They memorize the training data and they have high variance. Therefore, these models are overfitted.
In this paper, Random Forest Classifier, Logistic Regression Classifier, Decision Tree Classifier, Linear Discriminant Analysis, K-Nearest Neighbor Classifier and SVC methods are compared to predict mobile phone price class. The dataset is evaluated to verify that there are no missing values. In addition, standard scaling is utilized to dataset to obtain more convenient data for learning algorithms. Scaling is particularly important step for distance based classifiers. Feature elimination, which reduce the number of inputs, is another significant process to decrease computational cost. Feature scoring is an important method for selecting the best features. Moreover, determining feature selection method is dependent to features and label of the dataset. We used Mutual Information and ANOVA methods because of the dataset used in this study has mostly numerical data and categorical label. It is seen that ANOVA method having nearly same accuracy with Mutual Information method is more convenient for this dataset. Finally, accuracies of learning algorithms are increased with parameter optimization. SVC classifier has the highest test accuracy compared to other models. Furthermore, there is no overfitting and underfitting problem for SVC method as shown in the Figure 7. When obtained results are compared, it can be said that not every method fits every dataset. The right method should be determined to achieve high performance.

REFERENCES