Siyu Wang
DATA 606
Deliverable-3 Report

Mobile Price Range Prediction

Data Preprocessing

Before I do the modeling for my project, I have to complete the data preprocessing. First, I need to complete the data preprocessing to prepare for modeling, because each feature has a different range. I need Standardize features by removing the mean and scaling to unit variance. "standardScale" makes the processed data conform to the standard normal distribution, ie the mean is 0 and the standard deviation is 1. This can speed up the speed of gradient descent to find the optimal solution and improve the accuracy for the Non-probabilistic model. Secondly, I divided the data into a training set and a test set. 80% of the data is used as the training set, and 20% of the data is used as the test set.

Models

Nature of the Problem

As Output labels shows, 0 means low cost, 1 means medium cost, 2 means high cost and 3 means very high cost. We can be sure that this is a multiclass classification problem. A sample belongs to and belongs to only one of multiple classes, a sample can only belong to one class, and different classes are mutually exclusive. Based on the nature of the problem, I chose Logistic Regression, Neural Network and Random Forest to predict this problem. For machine learning models, I used the sklearn package. Because this is a multi-classification problem, I used Mean Accuracy as the metric to evaluate the model.

Logistic Regression

I firstly use Logistic Regression. Logistic Regression is mainly used for binary classification. The algorithm uses the one-vs-rest (OvR) scheme when facing multiclass case. One-Vs-All's idea is to turn a multi-class problem into multiple two-class problems. The idea of the change is just as the method name describes, choose one of the categories to be positive and make all other categories to be negative. Its advantages are high efficiency and fast speed, and it can train as many classifiers as there are categories. But its shortcomings are also obvious. Logistic regression belongs to the generalized linear model, so it does not perform well on linear inseparable problems. Then, I use the confusion matrix to show the results. Confusion matrix is a situation analysis table that summarizes the prediction results of the classification model in machine learning. The records in the data set are summarized in a matrix form according to two criteria: the True Labels and the predicted labels by the classification model. Where the rows of the matrix represent the true values and the columns of the matrix represent the predicted values; the mean accuracy of the model in the training set is 86.75%, and the mean accuracy in the test set is 83.25%. From the confusion

matrix, I found that the prediction error only occurred in adjacent labels. In other words, the model never predicted high cost as low cost.

Neural Network

Next, I used a neural network and wanted to obtain higher accuracy through its powerful non-linear capabilities. Neural Network has the Ability to learn and build models of nonlinear complex relationships. In addition, the neural network can be generalized. After learning from the initial input and its relationship, it can also infer the unknown relationship from the unknown data, so that the model can generalize and predict the unknown data. The disadvantage is that the neural network requires a lot of training data. After hyperparameter tuning, I chose Relu as Activation function and Adam as Optimization Algorithm. From the results, the overall accuracy has been improved. The accuracy of the training set reached 100% because of the powerful memory of the neural network. In the test set, accuracy is as high as 92%. It can also be seen from the confusion matrix that the neural network performs better than logistic regression, especially in medium and high cost.

Random Forest

Lastly, I choose to use random forest. Random forests can handle high-dimensional data without dimensionality reduction or feature selection. In addition, it is not easy to overfit. The most important thing is that it can judge the importance of features. The disadvantage is that it does not perform well in noisy data sets. Ram is the most important feature, and its importance is much higher than other features that are given by Random forests. Battery power is another important feature. From the plot, mobile feature like 3g, 4g, wifi, Bluetooth doesn't influence price that much. And this result is same with my exception, because from my life experience, both cheap and expensive mobile phones have these configurations. From the Result, the Mean Accuracy for the Train Data is 98.75% and Mean Accuracy for the Test Data is 86.75%. Its performance is better than logistic regression but worse than neural network.

Next Step

In this part, I have basically finished what the part IV needs, so the next step I may focus on the visualization of the predict outcomes.