REVENUE ESTIMATES BASED ON SPECIFIC FEATURES OF APPLE

Kadir Gökdeniz-Mehmet Bayram Alpay

Computer Enginnering Department

Ankara University

Ankara,Turkey

{20290344,20290310}@ankara.edu.tr

Abstract -Figuring out an interaction in terms of economic between features of products and price is a practical application to predict for suitable investments. We recommend knowledge-based decisions rather than an intutive action plan. Statistics have been added to our article to examine the price changes with the data that the Apple company has had since its establishment. In this study, we have a hyptosesis that some features cause more income and other features may cause less income for Apple. Therefore, investors may make a decion in more guaranteed way.According to result, we got approximately 98-99% accuracy rate.

Keywords-Apple products, investment, economic, predictions.

1.Introduction

All over the world, Apple is a well-known technological company. AI programs are always evolving, and Apple needs more work on improving qualifications. In addition to these, investors want to guarantee their investments. Therefore, the investor is looking for an action plan for Apple products. Our main purpose in this study is to increase awareness of Apple's existing products on the basis of certain categories and to reveal its economic perspective. Most of the old studies are economic based studies and the main purposes are to guide entrepreneurs. Apple's data been stored since the 1980s and data is increasing day by day. In this study, we wanted to present a new study about Apple products and increase the accuracy of prediction compared to previous studies.

In this study, we particularly focus on analyzing data with volume , date and can be profitable or not. To the end,the result is calculated our by using confusion matrix.

2. Material and Method

2.1. Dataset

A model is proposed using data collected from Apple revenue from 1980 to 2022. Apple revenue dataset contains 10559 different data covering 42 years in total based on short time intervals. This data is calculated by collecting data from all over the world, not Apple company in any region. In this study, the data is preprocessed by reasonable use of numpy, pandas, and matplotlib packages. The general characteristics of the dataset are shown below.

	open	high	low	close	volume	profit or not
count	10559.000000	10559.000000	10559.000000	10559.000000	1.055900e+04	10559.00000
mean	15.308827	15.483090	15.139373	15.318012	3.263275e+08	0.476466
std	33.973113	34.385796	33.580974	34.000790	3.201156e+08	0.499469
min	0.038800	0.038800	0.038400	0.038400	1.001504e+06	0.000000
25%	0.236800	0.242500	0.231650	0.236900	1.219792e+08	0.000000
50%	0.400200	0.406600	0.393000	0.399300	2.159780e+08	0.000000
75%	13.652700	13.783300	13.504000	13.647150	4.074518e+08	1.000000
max	181.877900	182.186600	178.382400	181.260500	2.147484e+09	1.000000

Figure 1. Describe data set

2.2. Method

Logistics regression model is selected in order to analyze the economic situation that occurs with the amount of money a company receives from its customers in exchange for the sales of goods or services according to years. The logistic regression model allows to examine the influence of many independent variables $XX1$, …, $XXkk$ the dependent variable Y. The variable Y takes only two values. These values are coded as 1 and 0. A value of the variable Y equal to 1 means that a desired event occurs. Otherwise, when an adverse event occurs, this variable assumes a value of 0. Logistic regression uses a logistic function for the description, whose output values are in the range (0; 1) and which creates a curve in the formation of the letter S. Three stages of changing the value of the function can be distinguished: initially, up to a certain threshold, they practically do not change the probability, after reaching the threshold value the probability increases to one and stays at this level. Logistic regression model is as follows:

$$f(z)=\frac{e^{z}}{1+e^{z}}=\frac{1}{1+e^{-z}},\quad z\in R$$

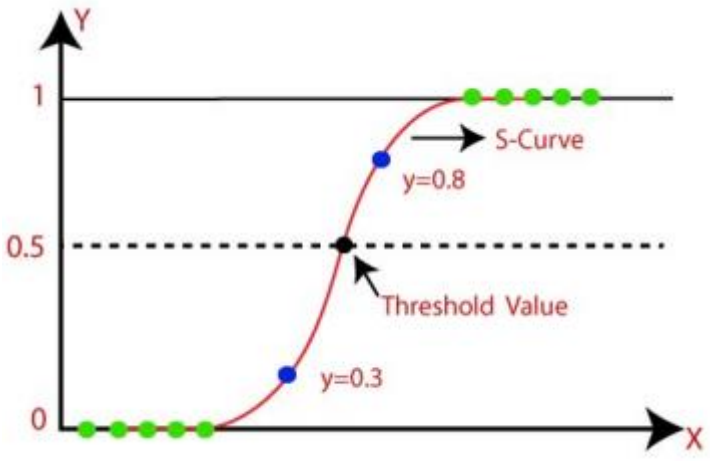

Figure 2. Logistic Regression model creates an S-like curve

After estimating the parameters of the logistic regression model, the theoretical values of the Y variable can be determined according to the standard estimation principle. For each value, the theoretical value will give a result between 0 and 1.

2.2. Handling Data and Comparison

According to the sample data, we used the isnull().sum() function to analyze the number of missing samples, respectively, and we saw that there were no missing or troublesome samples as a result of this function. Since there is no missing or troublesome data in the data set, we divided the data into two parts as 65% training and 35% test structure. After this part, we completely randomized the dataset with shuffle to maximize accuracy values, and oversampled using hyperparameters.

Because model accuracy is the focus of model training, different models are used to compare and analyze the data such as precision Rate, accuracy rate to determine the optimal classification model.

Logistic regression model

According to the confusion matrix shown in Figure 3, on the training set sample, there are 3,585 samples with correct predictions, 10 samples with incorrect predictions, and 3268 would give the correct result.

	True	False
0	3585	6
1	4	3268

Figure 3. Confusion Matrix These data correspond to the image in figure 4 as a percentile.

Figure 4. Training dataset confusion matrix

The confusion matrix data that comes out as a percentage when we use it to estimate the 35% test data we initially reserved is shown in Figure 5. In this image, it is seen that the accuracy rate is more than 99%. The main point in accessing this data is to use hyperparameters and overlearning to occur.

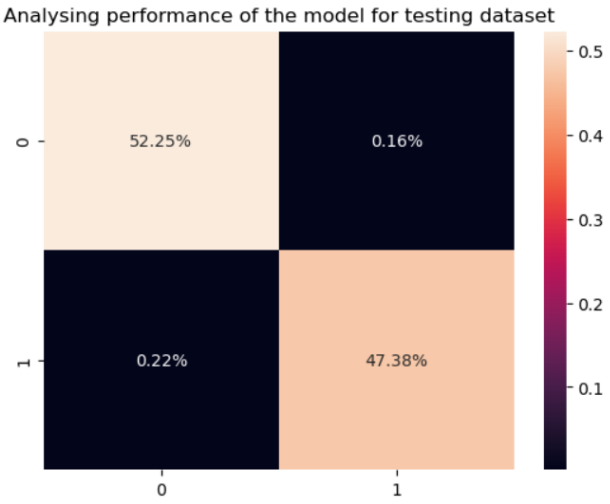

Figure 5. Logistic Regression testing dataset (comparing)

3.Results

The data used for training is divided into 'date', 'profit or not' and 'volume' headings. The title used for classification is just 'profit or not'. The results we reached in our project are stored in a data structure called confusion matrix. The data in this data structure gives percentage results proportionally.

In supervised machine learning, it is important to train an estimator on balanced data so the model is equally informed on all classes. Setting weights is estimator specific. Many Scikit-Learn classifiers have a class_weights parameter that can be set to ‘balance’ or given a custom dictionary to declare how to rank the importance of imbalanced data. In this method, it is similar to oversampling. Using weights, we can force as estimator to learn based on more or less importance (‘weight’) given to a particular class. Balanced class weights can be automatically calculated within the sample weight function. Set class_weight = 'balanced' to automatically adjust weights inversely proportional to class frequencies in the input data.

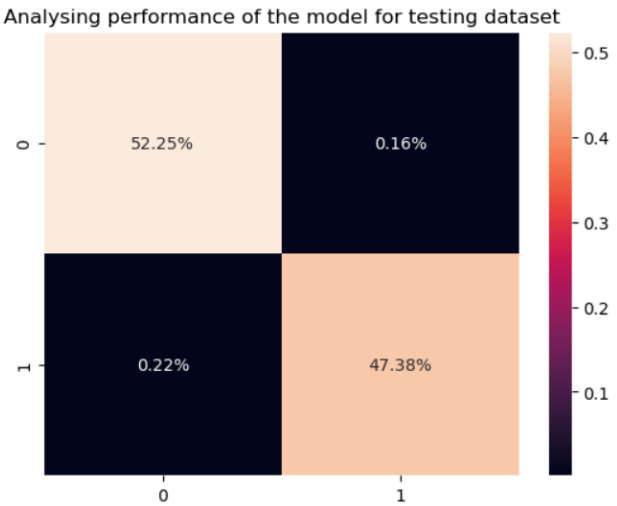

Figure 3.1

Sensitivity (Sn) is the ratio between correctly classified effectors and all effectors:

$$Sn = \frac{TP}{TP + FN} \tag{12}$$

Specificity (Sp) is the ratio between correctly classified non-effectors and all non-effectors:

$$Sp = \frac{TN}{TN + FP} \tag{13}$$

Accuracy (Acc) is the ratio between correctly classified non-effectors and effectors and all samples:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{14}$$

Figure 3.2

For the matrix shown in the Figure 3.1., the accuracy rate is the sum of the values in its major diagonal then divide by one all cells that is (52.25.+47.38) / (52.25+0.16+ 0.22+ 47.38) ≈00.9963. Sensitivity ratio is the value in the first column of the first row divided by the sum of the first columns in the first and second rows that is, 55.25/ (52.25+0.22) ≈0.9960. The specificity ratio is the quotient of the data in the second column of the second row by the sum of the data in the second row of the second column and the data in the second column of the first row that is 47.38/ (47.38+0.16) ≈0.9966. They can also be formulated as in Figure 3.2.

True negatives are represented as TN. True negatives are where actually investing is wrong and algorithm categories it the same. False positives are represented as FP. False positives are where actually investing is wrong and algorithm categories that investing is true. False negatives are represented as FN. False negatives are where actually investing is true but algorithm categories that investing is false.True positives are represted are as TP. True positives are where actually investing is true and algorithm categories is true.

In old studies, accuracy rate is approximately 80%. These old studies has only logistic regresion approch. In this study,some hyperparameters are used and data are shuffled to get more certain results. These details used to increase accuracy rate from 80% to 99%. Addition to these, the parameter of random state is used to get permanent randomize state. Figure3.3 displays old study about Apple products.

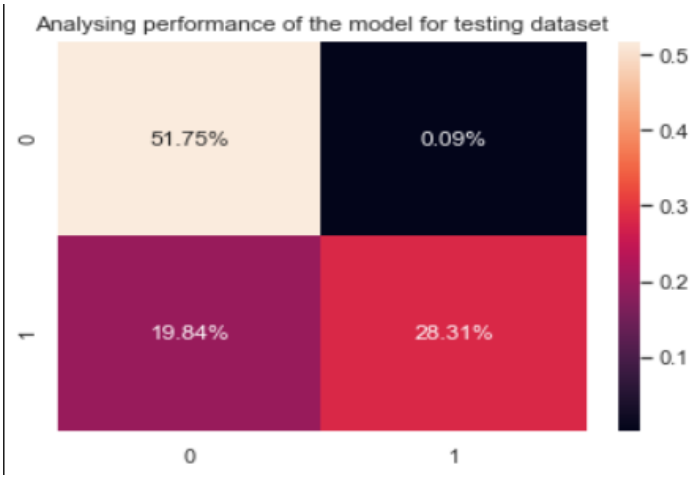

Figure 3.3

Some Categories about an Old Study

Accuracy ≈ 0.8006

Sensivity ≈ 0.7249

Specificity ≈0. 9968

4.Conclusion

As can be seen from figure 3.3, the accuracy rate has been increased from 80% to 99%. The sensitivity ratio was increased from 72% to 99%. These two are the two greatest achievements of the study. In these study, hyperparameters are used, data are shuffled and to provide permanent state with random state.These cause more remarkable our project.

There are some prons and cons in these project. First, accuracy is more higher than the others. Second, sensivity is also higher. These means that, our project is classifies more true result. However, our project has less specificity than the others. Therefore,the study classifies more true negatives.

In this study, some categories are used. In future studies, other categories will be used such as voice, weight e.g.

Logistic regression is generally used to evaluate categorical and numerical data. Therefore, it may be more appropriate to use knn-style algorithms, especially for data in string structure. In addition, the dependent variable, namely the result, must be of two types (binary). If these conditions are not met, the use of this structure will generate errors or lead to unexpected results. Logistic regression is an example of supervised learning. Its use in unsupervised learning poses a logical problem. Apart from these, the results obtained will not be satisfactory if excessive learning is not done with the class_weight = 'balanced' structure.

6.References

1. Koray Açıcı, Tunç Aşuroglu, Çağatay Berke Edaş, and Hasan Oğul, "T4SS Effector Protein Prediction with Deep Learning," Computer Science, vol 4(1), 45, 2019 https://doi.org/10.3390/data4010045

2. Koray Açıcı, Hasan Oğul, "Inferring Microarray Relevance By Enrichment Of Chemotherapy Resistance-Based MicroRNA Sets," Computer Science, vol ,2015

3. Yuzhen Wang, Jingqiao Qin, "Analysis of financial product purchases based on logistic regression," Information Engineering, Conf. Ser. 1848 012164, 2021 DOI 10.1088/1742-6596/1848/1/012164

4. Kevser Şahinbaş "Price Prediction Model for Restaurants In Istanbul By Using Machine Learning Algorithms," Ekonomi İşletme ve Maliye Araştırmaları Dergisi, vol 4 (2), 159-171, 2022 DOI: 10.38009/ekimad.1148216

5. Agnieszka Strzelecka a, Agnieszka Kurdyś-Kujawskaa, Danuta Zawadzkaa, "Application of logistic regression models to assess household financial decisions regarding debt", Economic Science, 176 (2020) 3418–3427