MICROSOFT STOCK QUOTES DEPENDENCY ANALYSIS

Janez Bucik
Data mining study class
University of Nova Gorica
Vipavska 13, SI-5000 Nova Gorica, Slovenia
Tel: +386 40425016
e-mail: janez.bucik@b-s.si

ABSTRACT

This paper presents Microsoft stock quotes dependency analysis. In this data mining task, I tried to find out some correlations between Microsoft corporation stock quotes and stock quotes of some other worldwide known information technology oriented companies. The main goal was, to figure out any dependencies between stock quotes and possibly represent them using existing data models in Weka software. I evaluated the models using "10-fold cross validation" and tested their performance on a separate example. The paper covers: data acquiring, data formatting, regression models and gained results during this work.

1 INTRODUCTION

The aim of this analysis is, to find any new interesting lawfulness between different stock quote movements. Stock quotes values change daily on the stock market. Normally the stock quote rises as the company's sales revenue and its incomings rise and vice versa. If the company is not doing well their stock quotes on the market would probably lower. Despite this general rule, stock quotes may rise or lower from many other reasons. The intuition I got when looking at daily stock quote value was, that stock quote value of one company depends on stock quote value of other companies. This idea may seem a nonsense at first sight, but it should be kept in mind, that if company's businesses and activities are tightly related and supposing that they share the same trade market, there should also be a correlation between their stock values.

Ideal result of this study would be an appropriate model, which would foretell chosen stock quote value on the basis of other company's stock values with sufficient certainty.

In this study I selected Microsoft corporation stock quotes as the target variable. The stock quotes of six other computer technologies companies represent attributes on which Microsoft's stocks may depend.

First the appropriate data will be collected from free internet data source [5]. Afterwards I will try to preprocess data, to get a valid data format for Weka software. Once data is loaded in Weka, suitable regression models will be built, in order to gain best foretell results. The paper will conclude a result representation and possible further directives.

2 DATA DESCRIPTION

The data I used derives from free stock quote rates [5]. The dataset consists of 7 stock quotes:

- MSFT – Microsoft Co.
- HPQ – Hewlett Packard Co.
- IBM – Intl Business Mach
- AAPLE – Apple Inc.
- AMD – Adv Micro Devices
- DELL – Dell Inc.
- ORCL – Oracle Co.

Each daily stock quote has the following values: open, high, low and close. For this research purposes I will only use "open" value of the ticker information, this will be the only attribute, my further work will base on. In general this value represents the start trading value for that day.

Data downloaded from the internet already had appropriate form, except the missing stock quote values for some days. The reason for this may be explained by the fact, that for some reason trading with that specific stock was not open for these days. As this case concerns, the selected period from 01.01.1990 to 20.04.2007 including 4349 instances in above selected stock quotes, fortunately had no missing values.

3 METHODS

Once data was loaded in Weka, the only problem seemed choosing the right regression methods. With help from our teaching assistant Petra Kralj and some base knowledge of machine learning and data mining procedures [1], two regression methods were selected for solving our regression problem.

- **M5P with regression tree**. This model should give us a detailed analysis of give data. As a result, we will have a quite large regression tree, that classifies our new input data. A regression tree will graphically represent the results.
- **LeastMedSq**. Using this model we will implement linear regression on our data and will be able to

foretell future stock quote value with a given certainty.

Both methods will be applied to imported dataset a few times in order to optimize model's parameters to get better results.

4 M5P APPLICATION

As described, I imported the data in Weka [2] and tried to launch the model a few times to tune the parameters. When running models, I used default "10-fold cross-validation". At first result looked ok as for regression tree output, but the number of leaves was about 250. To reduce the number of leaves and build a more representative model, I had to enlarge the number of instances in each leaf [3]. Default value in Weka is set to 4, what in my case resulted in exaggerated adaptation to given dataset, which is not appropriate for interpretation. After several attempts, the proper value for min. instances seemed to be around 300. This way I got an acceptable number of classification leaves that went reduced to 21. Figure 1 shows the resulting regression tree.

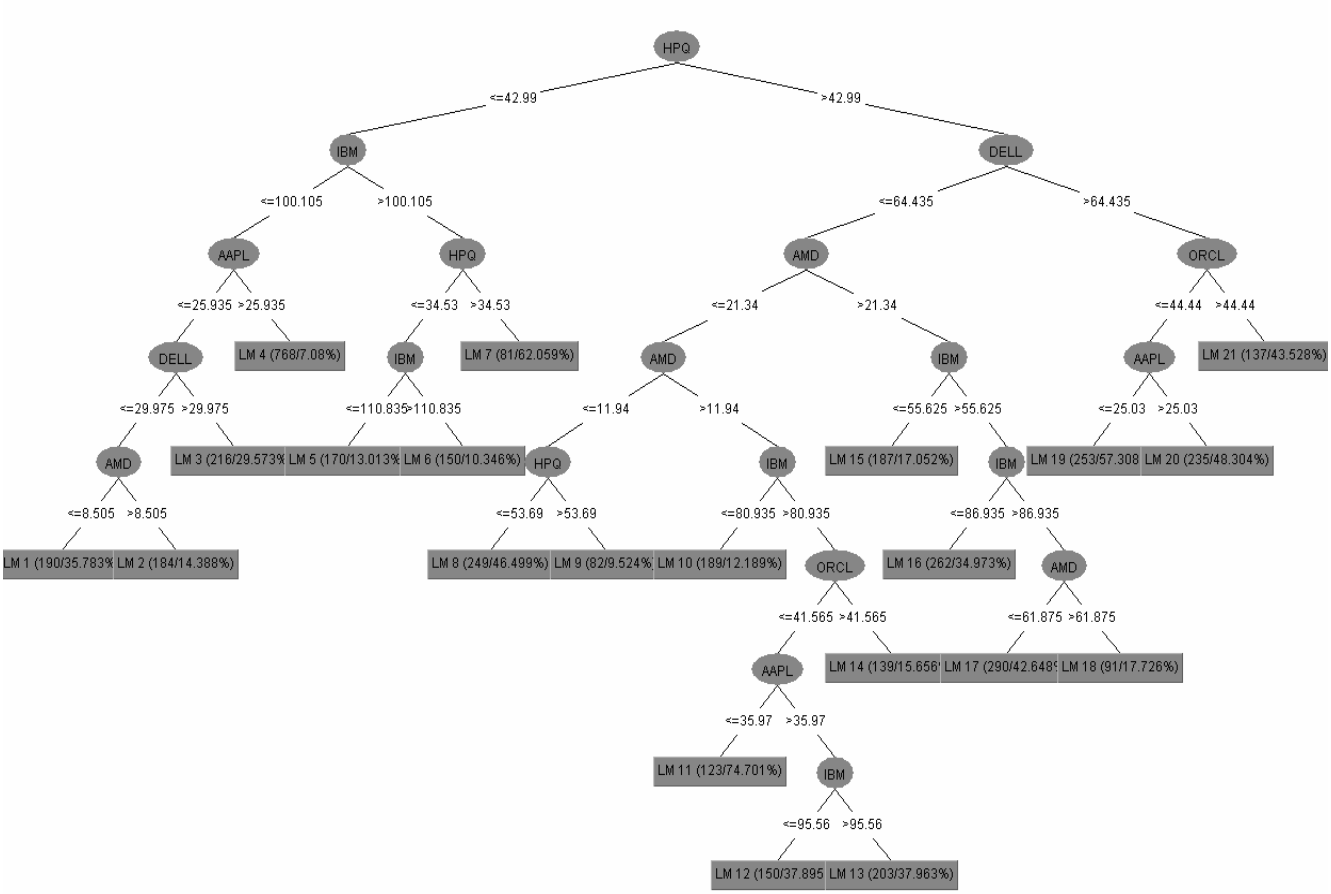

Figure 1: *M5P regression tree*

The regression tree leaves and their target value that actually build the model are show in Figure 2. Leaves are listed from the lowest possible 27.3499 $ to the highest 137.7953 $ Microsoft Co. stock quote value.

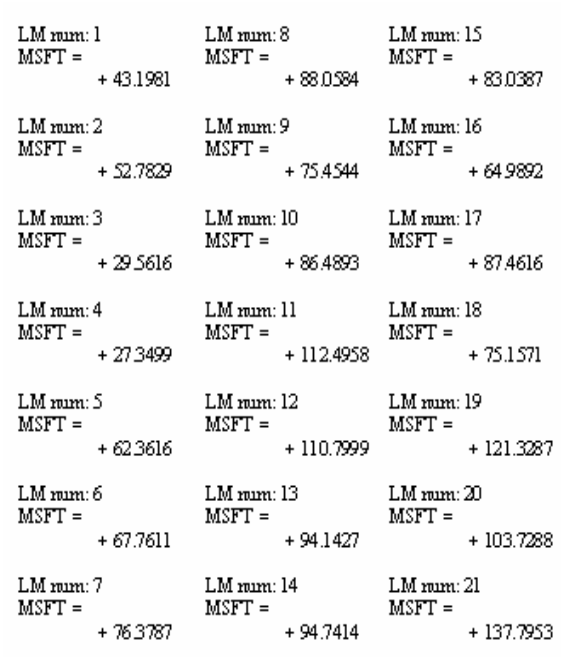

Figure 2: *M5P regression tree leaves*

Increasing the number of instances in leaves causes increases in error estimates like "Mean absolute error", but if we want a clear and representative model, we have to permit a little higher error estimates. The summary of M5P model evaluation is shown in the Figure 3.

Correlation coefficient	0.9247	
Mean absolute error	8.8477	
Root mean squared error	13.2724	
Relative absolute error	30.7963	%
Root relative squared error	38.1212	%
Total Number of Instances	4349	
Ignored Class Unknown Instances		1

Figure 3: *M5P model error estimates*

5 LEASTMEDSQ

Since there are very few parameters to set up, linear regression was much easier to apply to the selected dataset. In this case I used the default "samplesize" with value 4 [4]. As in M5P the "10-fold cross-validation" was used and the liner formula was calculated as show in the Figure 4.

```
MSFT =

      0.5533 * HPQ +
      0.3785 * IBM +
     -0.2854 * AAPL +
     -0.9289 * AMD +
      0.4937 * DELL +
      0.4543 * ORCL +
      3.4525
```

Figure 4: *Linear regression formula*

As in every model, there are some error estimates derived from data processing. In this LeastMedSq model they could not be much optimized and were calculated as shown in the Figure 5.

Correlation coefficient	0.7479	
Mean absolute error	18.3285	
Root mean squared error	23.3399	
Relative absolute error	63.7961	%
Root relative squared error	67.037	%
Total Number of Instances	4349	
Ignored Class Unknown Instances		1

Figure 5: *LeastMedSq error estimates*

6 EVALUATING THE MODELS

The models are now built, and a usual question that arises is "How good this models are?" and "Can I foretell future value of Microsoft Co. stock quotes with these concepts?" So, to test if these model are enough efficient, we will try to foretell a Microsoft Co. stock quote for one day, on the basis of other companies known stock quotes, with both selected models. Stock quote open values on the 25. April 2007, which were not included in the chosen training set for regression, were as show in the following table.

Table 1: *Stock quote open value on 25. April 2007*

MSFT	HPQ	IBM	AAPLE	AMD	DELL	ORCL
28.86	41.80	98.74	94.23	14.59	24.73	18.89

Using M5P model would foretell MSFT open value equaling to 27.3499 as result of LM num 4. Here is just one example shown, I have tested a few more day stock values, and on average the M5P model does not differ from the correct value for more then about 7 %.

Testing LestMedSq model with the same test example using LeastMedSq formula gives us 44.30, which exceeds the actual value for about 35 % and seems less reliable then M5P model what was to be expected.

7 CONCLUSION

Data modeling always brings several inaccuracies and can actually model data only with a certain certainty. It is impossible to build 100 % accurate models. For this study case both models seemed to be appropriate for regression modeling. Linear regression model gave a little worse result, but we have to keep in mind, that linear regression is not the best solution, when target variable is not linear dependent from attributes. Open values of Microsoft Co. in the training dataset, which varied from 21.59 $ to 178.94 $, seem to be a similar

case. That is why, LeastMedSq would probably not be the most suitable classification model to foretell future results.
Meanwhile M5P model did very good. Despite quite large error estimates, its result were quite impressive. This, I think, would confirm adequacy to use M5P model in similar cases, even though both chosen models are good, but as model evaluation concerns, M5P did twice better then LeastMedSq. Furthermore, it would be interesting to test some other regression models, that Weka has to offer and optimize their parameters precisely; it could result in even better outcome.

References:

[1] Witten, I. (et al.): Data Mining: Practial machine learning tools and techniques with Jave implementations. Second Edition, 2005, Morgan Kaufman
[2] Weka documentation: http://www.cs.waikato.ac.nz/~ml/weka/
[3] M5P overview: http://www.dbs.informatik.uni-muenchen.de/Lehre/KDD_Praktikum/weka/doc/weka/classifiers/trees/M5P.html
[4] LeastMedSq overview: http://www.dbs.informatik.uni-muenchen.de/Lehre/KDD_Praktikum/weka/doc/weka/classifiers/functions/LeastMedSq.html
[5] Data source: http://www.shareup.com/HSQuote_Stock_Quote_Downloader-download-2041.html