Capstone Project Report

Bruno Kiyoshi Ynumaru
ynumaru@gmail.com
Machine Learning Engineer Nanodegree
Udacity

November 6, 2020

Introduction

Time-series prediction of asset prices or asset price movements is one the most studied Deep-Learning applications in Finances.

Technical Analysis is a field in Finances in which an analyst makes predictions on future asset prices based on price charts. Technical Analysis fundamentals were postulated by American journalist Charles Dow, in the late XIX century, in what is known as the Dow Theory.

There are six main components to the Dow Theory, namely:

1. The Market Discounts Everything
2. There Are Three Primary Kinds of Market Trends
3. Primary Trends Have Three Phases
4. Indices Must Confirm Each Other
5. Volume Must Confirm the Trend
6. Trends Persist Until a Clear Reversal Occurs

The detailed explanation of each of these components is beyond the scope of this introduction.

The commonly used type of chart for analysis in this field is the candlestick chart, where the horizontal axis represents time, and the vertical axis represents asset price.

Figure 1Candlestick chart of EUR/USD currency pair on daily timeframe in MetaTrader 5 trading platform. (CC BY-SA 4.0)

Each time-step in a candlestick chart is called a candle, and it indicates four basic values: the open price, the closing price, and the maximum and minimum values a given asset is negotiated at in each timeframe.

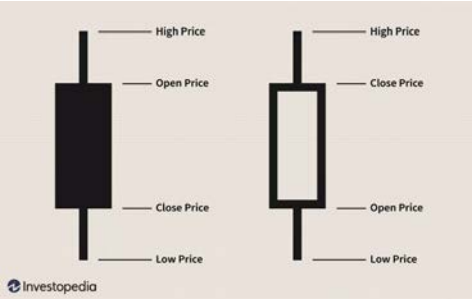

Figure 2Image by Julie Bang © Investopedia 2019 (available at https://www.investopedia.com/thmb/9nwzVl-16xZlP9epM3b2H7kdBSc=/6250x0/filters:no_upscale():max_bytes(150000):strip_icc():format(webp)/UnderstandingBasicCandlestickCharts-01_2-7114a9af472f4a2cb5cbe4878)

Classically, white or green candles indicate that the closing price is higher than the opening price, and the opposite is indicated by black or red color. Historically, some candlestick patterns have shown to indicate future movements of asset prices (they can be both indications of reversal and continuation movements). Some of these patterns are:

1. Bullish Engulfing
2. Hammer
3. Bullish Harami
4. Bearish Engulfing
5. Falling Star
6. Bearish Harami
7. Doji
8. Three Line Strike
9. Two Black Gapping
10. Three Black Crows
11. Evening Star
12. Abandoned Baby

These patterns can be visualized in Figure 3.

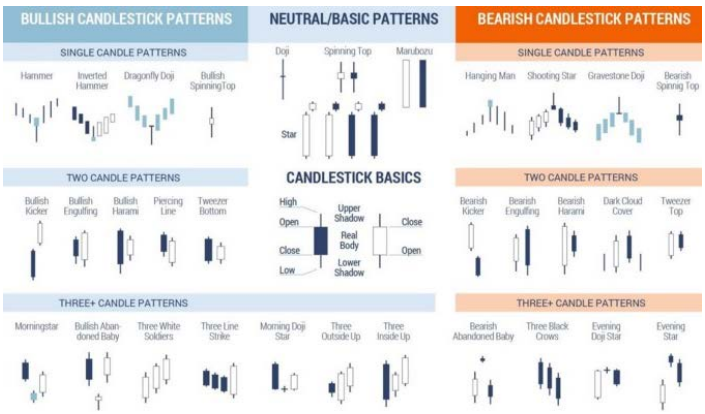

*Figure 3Visualization of some of the main candlestick patterns used in Technical Analysis. (Available at https://miro.medium.com/max/700/1*OpahFyAd6nHkdVtO59DX9w.jpeg)*

Detailed explanation of how these patterns work and what they predict is also beyond the scope of this document.

Some other features are commonly used in technical analysis, such as:

1. RSI – Relative strength index
2. Moving averages (linear or exponential)
3. Bollinger bands
4. True range
5. Financial volume (in shares or in currency)

Problem definition

Financial market forecasting can be both a regression and a classification problem. When one tries to predict the price a given asset will be negotiated at in the future, it's a regression problem. Simply trying to foresee whether a price is going to rise, fall, or continue the same becomes a classification problem. Similar methods can be used to predict the scores or movements of indexes such as the S&P 500 – and related ETF's if existent.

In this work, we will be trying to predict both prices and market movements for a Brazilian company's stock. WEG operates worldwide in the electric engineering, power and automation technology areas, and is one of Brazil's most prominent companies in 2020.

In order to perform these predictions, data from one other asset will be used: BOVA11 is an ETF that follows the IBOVESPA index. In theory, it is actually WEGE3's prices that affect such this index, but our hypothesis is that we can also use the index as an indicator of how the overall market is going to behave in the future.

Dataset analysis

Two datasets, one for each asset, were exported from the *"Nelogica Algotools"* software, though stock prices data are publicly available from several sources. The same software exported some technical analysis features. Namely: RSI, True Range and Trading Volume (both in BRL and in number of negotiated shares).

The following figures present the originally obtained data in plots:

Figure 4 WEGE3 OCHL prices history

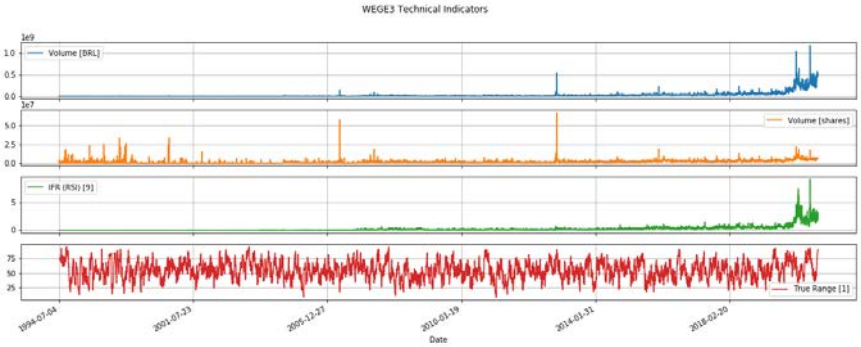

Figure 5 WEGE3 technical indicators generated in Nelogica Algotools

Date	Open	High	Low	Close	Volume [BRL]	Volume [shares]	IFR (RSI) [9]	True Range [1]
2020-10-16	81.06	83.36	80.91	82.05	525025471.0	6355600	2.59	89.56
2020-10-15	79.64	81.28	78.10	80.77	490407232.0	6096300	3.18	88.37
2020-10-14	79.26	80.98	79.26	80.58	527948834.0	6579000	1.83	88.19
2020-10-13	77.88	79.47	77.35	79.15	473625847.0	6030100	2.57	86.84
2020-10-09	74.94	77.18	74.55	76.90	438389500.0	5745100	2.63	84.34
...	...	...	...	...	...	...	...	...
1994-07-08	0.04	0.04	0.04	0.04	152830.0	1666340	0.00	NaN
1994-07-07	0.04	0.04	0.04	0.04	393366.0	4495400	0.00	NaN
1994-07-06	0.03	0.04	0.03	0.04	6195.0	74360	0.00	NaN
1994-07-05	0.03	0.03	0.03	0.03	1925.0	23660	0.00	NaN
1994-07-04	0.03	0.03	0.03	0.03	267.0	3380	NaN	NaN

Figure 6 WEGE3`s first and last five rows of original data

	Open	High	Low	Close	Volume [BRL]	Volume [shares]	IFR (RSI) [9]	True Range [1]
count	5660.000000	5660.000000	5660.000000	5660.000000	5.660000e+03	5.660000e+03	5659.000000	5652.000000
mean	7.189864	7.308733	7.080466	7.190986	2.421622e+07	1.696467e+06	0.239435	54.764418
std	10.367506	10.575458	10.195485	10.388577	6.375940e+07	2.378154e+06	0.494385	15.186430
min	0.030000	0.030000	0.030000	0.030000	2.670000e+02	3.380000e+03	0.000000	8.820000
25%	0.745000	0.757500	0.745000	0.750000	3.395825e+05	3.627585e+05	0.010000	44.115000
50%	4.170000	4.230000	4.100000	4.160000	4.826972e+06	1.147172e+06	0.110000	54.555000
75%	10.120000	10.302500	9.942500	10.092500	2.168816e+07	2.295058e+06	0.260000	65.400000
max	81.060000	83.360000	80.910000	82.050000	1.172040e+09	6.801033e+07	9.000000	95.270000

Figure 7 WEGE3`s original data description

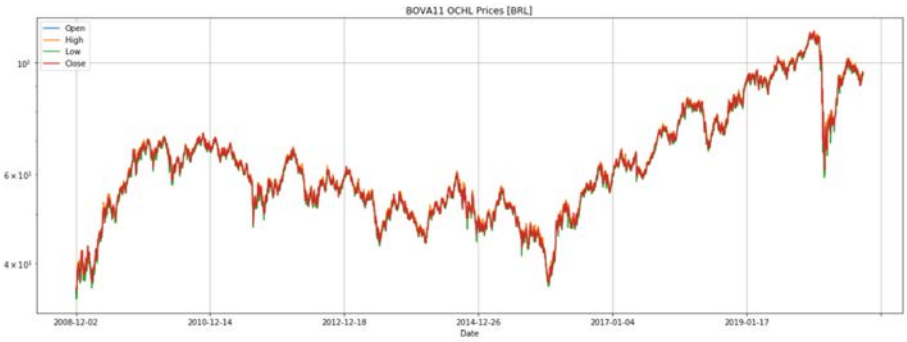

Figure 8 BOVA11`s OCHL prices history

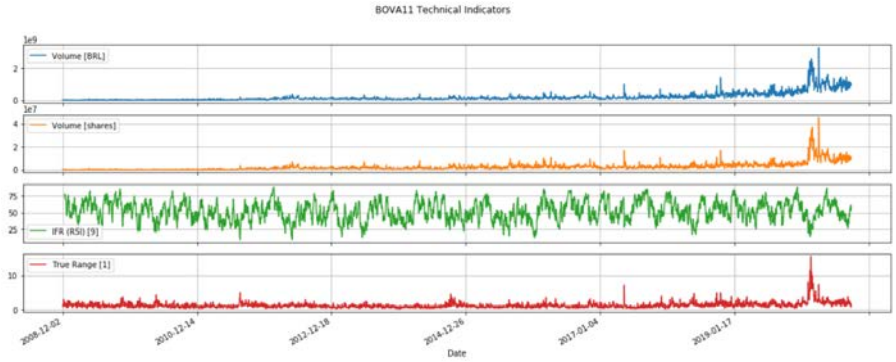

Figure 9 BOVA11 technical indicators generated in Nelogica Algotools

Date	Open	High	Low	Close	Volume [BRL]	Volume [shares]	IFR (RSI) [9]	True Range [1]
2020-10-16	95.12	95.45	94.54	94.54	9.673009e+08	10188511	54.42	0.91
2020-10-15	94.36	95.80	94.05	95.34	9.081088e+08	9544784	59.67	1.75
2020-10-14	95.01	95.86	95.01	95.59	1.092614e+09	11431930	61.32	1.06
2020-10-13	94.10	95.35	93.65	94.80	9.415492e+08	9960394	58.07	1.70
2020-10-09	94.13	94.99	93.50	93.65	8.611634e+08	9149800	52.96	1.49
...	...	...	...	...	...	...	...	...
2008-12-08	37.35	38.46	36.95	38.46	3.857295e+06	103100	NaN	3.11
2008-12-05	34.51	35.35	34.11	35.35	6.671196e+06	193700	NaN	1.59
2008-12-04	35.80	36.00	35.15	35.70	4.554168e+06	128400	NaN	0.85
2008-12-03	34.73	35.40	33.90	35.31	1.148272e+07	330100	NaN	1.50
2008-12-02	35.11	35.81	34.91	35.39	2.681997e+07	759900	NaN	NaN

Figure 10 First and last rows of data of BOVA11 data

	Open	High	Low	Close	Volume [BRL]	Volume [shares]	IFR (RSI) [9]	True Range [1]
count	2936.000000	2936.000000	2936.000000	2936.000000	2.936000e+03	2.936000e+03	2928.000000	2935.000000
mean	64.107006	64.664939	63.452142	64.074029	1.979881e+08	2.727684e+06	52.383282	1.344508
std	16.816599	16.888117	16.696268	16.793429	2.800640e+08	3.319674e+06	14.718881	0.875526
min	34.510000	35.350000	33.900000	35.310000	1.816440e+05	4.400000e+03	9.910000	0.170000
25%	52.110000	52.520000	51.500000	52.017500	4.614790e+07	8.376700e+05	41.467500	0.860000
50%	60.140000	60.685000	59.610000	60.150000	1.036209e+08	1.851050e+06	52.650000	1.160000
75%	70.717500	71.300000	70.012500	70.720000	2.164058e+08	3.348998e+06	63.395000	1.580000
max	115.210000	115.260000	113.710000	115.210000	3.295958e+09	4.589951e+07	88.730000	15.760000

Figure 11 BOVA11 data description

Since we will also be trying to predict the direction of price movements, it`s a good ideia to assess how frequently WEGE3`s prices went up or down. Price is considered to have gone up when the closing price at a given day (D+1) is greater than the closing price from the previous day (D).

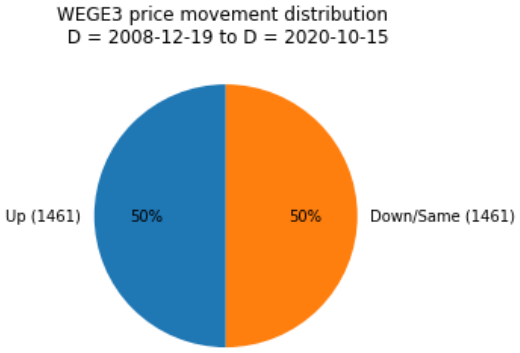

Figure 12 Price movement distribution for WEGE3

Code implementation

All code was implemented on AWS SageMaker platform. The Notebook instance type used was ‘ml.t2.medium’, which counts on 2 virtual CPU and 4GiB of RAM.

Code was implemented in two main Jupyter notebooks, namely “feature_engineering.ipynb” and “capstone_project.ipynb”, plus two Python scripts (“train/train.py” and “train/model.py”) for the regression model and two other (“train_classifier/train.py” and “train_classifier/model.py”) for the classification model.

“feature_engineering.ipynb”

This notebook runs on a Python 3.7 kernel and relies on the following Python packages: “os”, “pandas”, “matplotlib”, “numpy” and “datetime”.

The code in this notebook is responsible for taking the price data from both our datasets and engineering new features into them. The features that were engineered in this notebook could also have been created using a python package called “TA-Lib” - which stands for “Technical analysis Library”. Some of the motivations such library was not applied in this work are:

1. “TA-Lib”, despite being able to detect candlestick patterns, does it in a binary way. In other words, it merely indicates the detection (or lack thereof) the candlestick pattern. I wished to indicate not only the presence of these patterns, but also their strengths when possible.
2. The detection of some candlestick patterns should depend on whether price trends are in upward or downward movements. Some tests I ran using “TA-Lib” drove me to believe the package makes no such distinction. For example, the “Bearish Harami” is a candlestick pattern which indicates that a price movement is about to switch from a bullish to a bearish trend. Therefore, it makes no sense to detect such pattern during a bearish movement.
3. Author`s will to implement candlestick pattern detection from sctrach.

“capstone_project.ipynb”

This notebook runs on a Python 3.6 kernel and relies on the following Python packages: “os”, “sklearn”, “pandas”, “datetime”, “numpy”, “matplotlib”, “torch”, “sagemaker” and “boto3”.

The “capstone_project” file is the main body of the Deep Learning analysis that was made. It starts by transforming the engineered dataset into a format that is fit for supervised learning applications, in which input and output data are in the same pandas DataFrame rows. Both datasets were merged in a way we were only left for WEGE3 data which had a correspondent date in the BOVA11 dataset.

Test and training datasets were split so that we had two months of testing days, and all previous dates were used as training set.

All data was scaled using scikit-learn’s MinMaxScaler. The feature ranges applied were minimum and maximum values from the training, which were then applied to scale the testing set.

Regression model

For the regression task, an LSTM model was applied. This model contained one LSTM layer with 313 neurons. This architecture was the result of a hyperparameter tuning job done inside SageMaker.

Figure 13 and Figure 14 show the model’s fit to the training data after 1000 epochs.

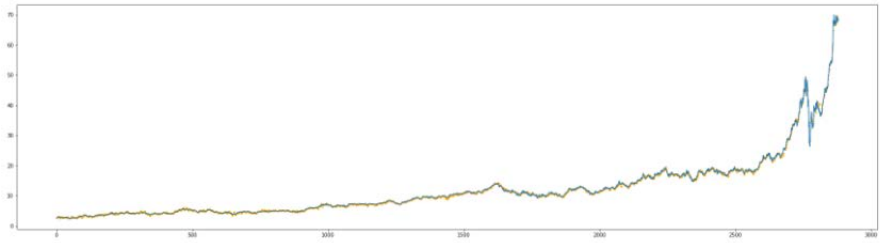

Figure 13 LSTM fit to training set

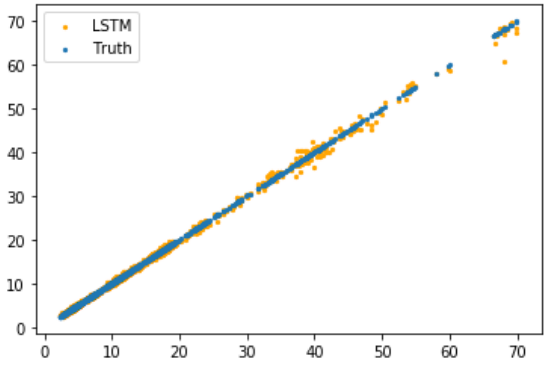

Figure 14 LSTM fit to training set

Classification model

For the regression task, an LSTM model was applied. This model contained one LSTM layer with 118 neurons. This architecture was also the result of a hyperparameter tuning job done inside SageMaker.

Results

Figure 15 shows how the model predicted prices (orange) versus the actual prices found in the dataset. The means squared error in this case was 8.7.

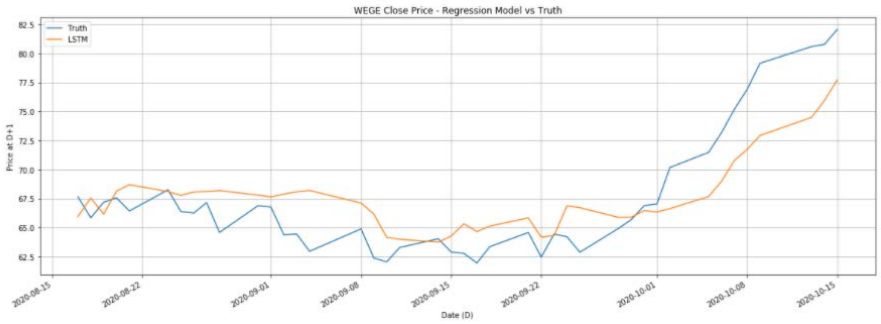

Figure 15 Regression model vs true values

The LSTM classifier had an accuracy of 0.643, a recall of 0.800 and a precision of 0.667.
The results can be visualized in the figures below. Figure 16 shows in green the sections of the graph for which the model had predicted an upward movement, whilst red sections indicate the model pre dicted a downwards movement.

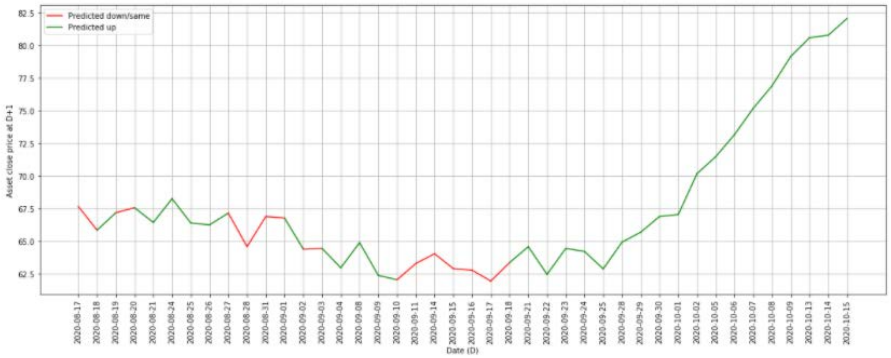

Figure 16 Predicted movements

Comentado [BY1]:

Figure 17 presents in blue the movements the model correctly predicted and in red the movements that were incorrectly predicted.

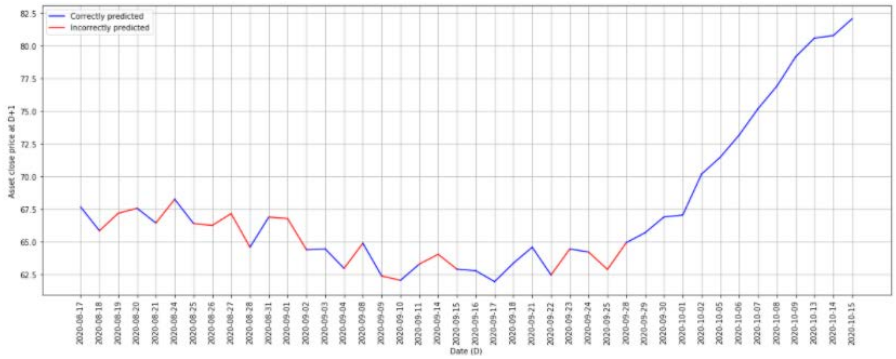

Figure 17 Correct and incorrect movement predictions

Figure 18 presents how many true positives, true negatives, false positives and false negatives our classificator resulted in. By “positive”, the plot represents upwards movements.

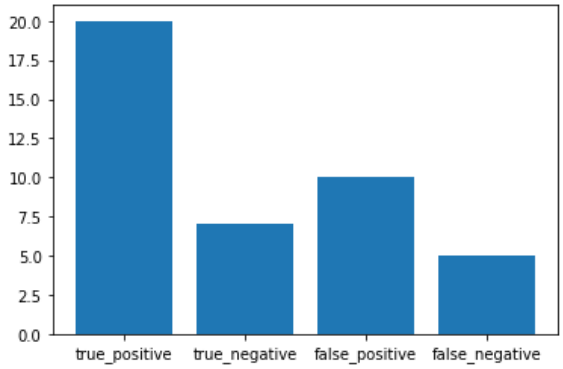

Figure 18 Classification results

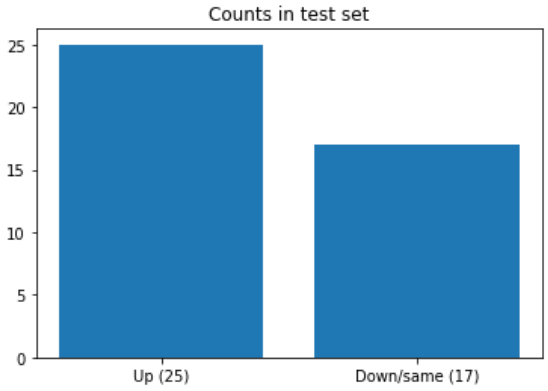

Figure 19 Movement counts in test dataset

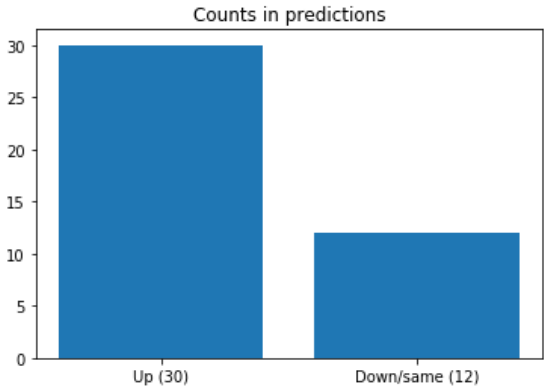

Figure 20 Predicted movement counts

Conclusions and future work

Two LSTMs were trained and applied to a set of price data and technical indicators. The regression model, which aimed to predict future prices of WEGE3 stocks resulted in a mean squared error of 8.7, which in the authors opinion is not a good enough result so that such model could be used in real life. The graph plotted from predicted prices shows trends in prices are somewhat accurately detected by the deployed model.

The classification model correctly predicted price movements more than 60% of times, which is a good result compared to the baseline 55% achieved by (David M. Q. Nelson, 2017).

If this work were to be continued, some new goals would be:

1. Experimenting with other models such as GRU's.
2. Experimenting different training and testing sets lengths.
3. Trying the experiments on other assets.
4. Calculating the financial return an investor blindly following the models would have.

References

David M. Q. Nelson, A. C. (2017). *Stock Market's Price Movement Prediction With.* Belo Horizonte – MG – Brazil: Universidade Federal de Minas Gerais.