Ibovespa forecasting using neural networks

Machine Learning Engineer Nanodegree

Capstone Proposal

Adelmo M. A. Filho
January 4th, 2020

Proposal

Domain Background

Bovespa Index (Ibovespa) is one of the most important benchmark indices traded on the B3, stock exchange located in São Paulo, Brazil. Ibovespa takes into account around 80 stocks that comprehend brazilian companies from multiple sectors (financial, mining, oil & gas, electric utilities). These stocks are reviewed every four months, when their participation percentage can be modified (FINKLER, 2017).

Described as an indicator of the average performance of the most tradeable and representative assets of the Brazilian stock market (FARIA, 2012), Ibovespa fluctuations tend to represent important aspects of brazilian economy, such as foreign investments, monetary policy decisions and political issues.

Problem Statement

For this project, a time series regression to predict the closing value for Bovespa index for the next trading day is proposed.

In general terms, Bovespa index closing value can be defined as a function of its previous values (endogenous variables) and independent (exogenous) variables, for example, calendar variables (weekday, month), stock prices, dollar exchange.

Models of this nature provide useful support to decisions and allow simulations of different scenarios and the understanding of variables importance for the Bovespa index closing value.

Datasets and Inputs

The datasets are provided by the python package Yahooquery which works as a wrapper for an unofficial Yahoo Finance API. Data used on this project was obtained for free, there was no need for a Yahoo Finance premium subscription.

The `history` method from `Ticker` class of Yahooquery package allows to retrieve daily data about stock markets. The following table shows a sample of historical data for the Bovespa Index.

symbol	date	open	close	low	high	volume
ˆBVSP	2020-04-08	76335.0	78625.0	76115.0	79058.0	10206300.0
ˆBVSP	2020-04-07	74078.0	76358.0	74078.0	79855.0	11286500.0
ˆBVSP	2020-04-06	69556.0	74073.0	69556.0	75260.0	9685400.0
ˆBVSP	2020-04-03	72241.0	69538.0	67802.0	72241.0	10411300.0

Not only Bovespa Index data is expected to be used on this project, but also historical data from the main stocks that represents its portfolio. The following table presents the main stocks that compose Bovespa Index and their global participation on the portfolio.

Ticker	Company	IBOVESPA Participation
ITUB4	Itaú Unibanco Holding S.A.	10,50%
BBDC4	Banco Bradesco S.A.	9,12%
VALE3	Vale S.A.	8,59%
PETR4	Petróleo Brasileiro S.A. - Petrobras	7,06%
PETR3	Petróleo Brasileiro S.A. - Petrobras	5,14%
ABEV3	Ambev S.A.	5,14%
BBAS3	Banco do Brasil S.A.	4,47%
B3SA3	B3 S.A. - Brasil, Bolsa, Balcão	4,15%
ITSA4	Itaúsa - Investimentos Itaú S.A.	3,86%

Solution Statement

The proposed solution for this project is to train and deploy a LSTM recurrent neural network combined with additional layers in order to create a complex neural network able to predict the Bovespa index closing value for the next trading day. These additional layers will have exogenous variables as input.

An additional neural network will be created in order to estimate if Bovespa index closing value delta between days is positive or negative. The delta sign from this additional model will also be used as an input variable.

The neural network architecture and hyper-parameters will be defined through grid-search techniques using an additional out-of-time validation dataset.

Benchmark Model

For this project, the benchmark model will be a simple moving average model with period equal to one, which means the prediction of Bovespa index closing value for the next trading day will be equal to the closing value of the index on the current day.

Evaluation Metrics

Model performance will be evaluated using an out-of-time sample (test dataset) of the last 3 months to estimate two metrics:

- Median absolute error regression loss: This metrics helps us to understand if the model predicts values with low error. The median calculation is insensitive to outliers, a good propriety in order to select a robust estimator.
- F1-score: Predictions should not only have low absolute error, it is important for the model to estimate correctly if the index value for the next trading day will increase or decrease. In order to achieve this understanding about a model, the sign of index value variation of one day

will be calculated for test dataset and predictions in order to calculate the F1-score.

It is important to notice that it is not expected good metric values for both metrics. Bovespa index closing values are highly influenced by events such as political decisions and news, which are out of scope for this project.

Project Design

Project design is divided into two perspectives: modeling and deployment.

Modeling

For modeling purposes, CRISP-DM Methodology (Figure 1) will be adopted. This methodology implies on a constant feedback between its stages and the understanding that data science processes are not linear. The main stages are broken down as follow.

- Data understanding: At this stage, it is planned to explore data distribution and visualizations in order to obtain insights about the modeling problem. Histograms, boxplots, scatterplots are examples of tools expected to be applied.
- Data preparation: Datasets will be joined and features will be created for posterior confirmation of their importance on the model. It is expected to work with stock market data and calendar variables to create multiple features, which combined with our target creates the modeling dataset.
- Data modeling: At this point, the modeling dataset will be split on training, validation and test datasets. Each one corresponds to an out-of-time sample from the modeling dataset. Recurrent neural networks will be created using `pytorch` and `AWS Sagemaker`, it is expected to test different layers to achieve the best architecture for this project.
- Evaluation: As describe previously, Median absolute error regression loss (MAE) and F1-score will be employed to select the best model. A low MAE metric value must also have a high F1-score in order for obtain a meaningful prediction.

Model deployment

This capstone project intends to create a web application where users can view predictions for the next days of Bovespa index closing value. To achieve such objective, the following architecture is proposed (Figure 2).

In details, every day a scheduler triggers a lambda function in order to collect new data from yahoo finance API. When new data is written at a S3 bucket, a event is triggered to a step-function that prepares input data at a lambda layer and starts a Sagemaker batch transform job to make a new prediction.

Predictions and raw data are displayed at a web application on the top of a EC2 instance, which users can access through Route S3 routing traffic service.

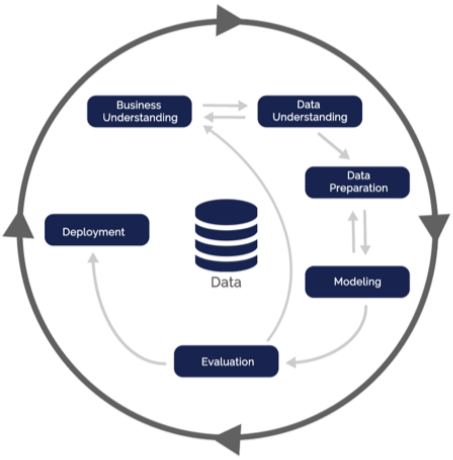

Figure 1: CRISP-DM Methodology

The focus of this architecture is the serverless approach of machine learning deployments. It is important to notice that eventual modifications and improvements on the model, can be easily added on this architecture, as well a model recalibration flow.

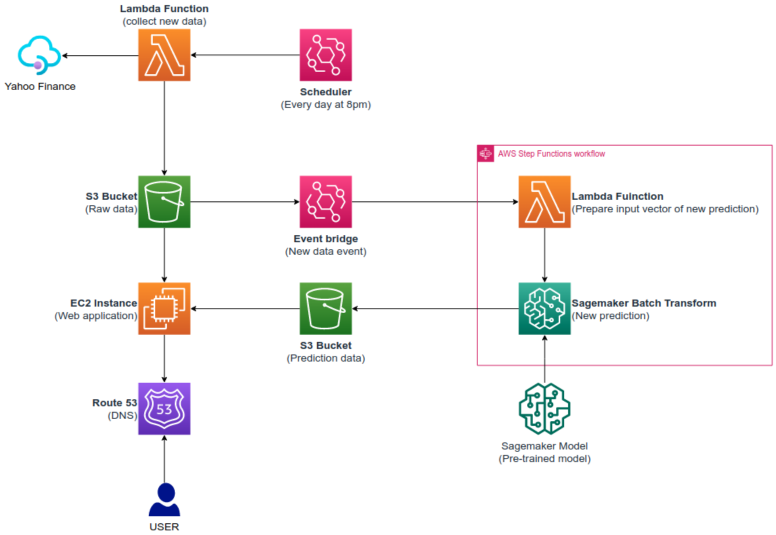

Figure 2: Model deployment architecture

References

- Faria, Elisangela Lopes. Uma metodologia para previsão do índice BOVESPA utilizando Mineração de Textos – Rio de Janeiro: UFRJ/COPPE, 2012.
- Finkler, Aline Cristiane. Aprendizagem de máquina aplicada à previsão dos movimentos do Ibovespa – Curitiba, 2017