Coffee Future Price Pediction Project: ICFFUT B3 Dataset

HarvardX - PH125.9x Data Science: Capstone Course

Mauricio Rabelo Soares

03 set 2022

Introduction

A Machine Learning for Factor Investing is a equity **investment strategies** that are built on **firm characteristics**, the goal of this approach is to determine the model that maps the time-t characteristics of firms to their future performance.[1] Factor investing is a subfield of a large discipline that encompasses asset allocation, quantitative trading and wealth management. Its premise is that differences in the returns of firms can be explained by the characteristics of these firms. In this project we use a similar approach but, instead a firm performance, our outcome is the 4/5 Arabica Coffee Futures contract performance traded at B3, which ticker is ICF.[2] The goal of this project is to create a decision trade system, which balanced accuracy, or F1 score, is above 50%, [3]using all the tools we have learn throughout the multi-part course in HarvardX's Data Science Professional Certificate series. For this challenge we going to use a dataset provided by the Nelogica and Clear[4]. The dataset begin in January 2003 and has **4435 decisions** applied to **1 asset** with **20 features**.

Methods

The methods section explains the process and techniques used in this project, in the first part, then explains the data cleaning used to extract and clean the data, in the second part. In the third part of this section we present the data exploration and visualization of the data to gain some insights. The fourth, and last part of this section, we show the modeling approach used in this project.

The process and techniques used

The decision trade system is similar to a **decision support system** (**DSS**) which is an interactive software-based system intended to help decision makers (future contract trader) compile useful information from a combination of raw data, documents, and personal knowledge, or business models to identify and solve problems and make decisions[5]. The future trader decisions could be **buy** the contract, **sell** or not get in a

[1] http://www.mlfactor.com/index.html

[2] https://www.b3.com.br/en_us/products-and-services/trading/commodities/product-sheet-8AE490C96D41D3A2016D46017EC97262.htm

[3] "The mean squared error is usually hard to interpret. It's not easy to map an error on returns into the impact on investment decisions. The hit ratio is a more intuitive indicator because it evaluates the proportion of correct guesses (and hence profitable investments). Obviously, it is not perfect: 55% of small gains can be mitigated by 45% of large losses. Nonetheless, it is a popular metric and moreover it corresponds to the usual accuracy measure often computed in binary classification exercises. Here, an accuracy of 0.542 is satisfactory. Even if any number above 50% may seem valuable, it must not be forgotten that transaction costs will curtail benefits. Hence, the benchmark threshold is probably at least at 52%."

[4] https://corretora.clear.com.br/plataformas/profit-trader-clear/

[5] https://en.wikipedia.org/wiki/Decision_support_system

position, what is the same to be **neutral** in a asset, this decisions will be our **outcome** $\mathbf{y} = y_t$, which is categorical.

Every trade decision taken, in a future market, occur in a certain time t, in a specific contract and organized marked, for a unique value. For this project the contract is the 4/5 Arabica Coffee Futures traded at B3, which ticker is ICF, and are available for march, may, july, september and december. The ICFFUT is the representative time-series of current contract and some of thus features will be our the dataset. The dimension of the feature matrix X is $T \times K$: where T are daily **observations** and each one of them has K **features**, **inputs**, or **predictors** which will serve as **independent** and **explanatory** variables.

The decision trade system build in this project are made using a similar process and techniques presented at (Coqueret and Guida 2020) and some approaches used by Renaissance Technologies.[6] The backtesting protocol propose by (Arnott, Harvey, and Markowitz 2018) are used as reference to evaluate the model.

The process begin with the download, then read the data and create some extensions using the Close price. After the data cleaning we explore the data to gain some insight trough visualization and selected table. This insights is the basis of the modelling approach of this project, that have 3 non linear algorithms: k-Nearest Neighbour Classification, Recursive Partitioning and Regression Trees and Classification and Regression with Random Forest

Data cleaning

For this project we going to use a subset of dataset provided by Nelogica and Clear, and for the sake of **reproducibility**, we will illustrate the concepts based on a single financial dataset available at https://github.com/mrabelosoares/Coffee-Future-Price-Prediction. This dataset comprises information on 1 contract listed at B3, which the time range starts in January 2003 and ends in August 2022. For each point in time, 19 **characteristics** describe the decision in the sample. The dataset are divide between raw data and extensions:

Raw data: Date; Asset; Price (Open, High, Low, Close); Volume; Decision; Open Interest

```
library(caret)
library(data.table)
library(fields)
library(tidyverse)
library(knitr)
library(kableExtra)
library(grid)
library(ggplot2)
library(lattice)
library(gridExtra)
library(readxl)
library(dplyr)
library(purrr)
library(zoo)
library(runner)
library(quantmod)
library(rpart)
library(randomForest)
library(xts)
library(TTR)
```

[6] https://www.youtube.com/watch?v=lji-jNsXmAM
Renaissance Technologies - Trading Strategies Revealed | A Documentary

```
##Data Clean

#create tempfile and download
dl <- tempfile()
download.file("https://github.com/mrabelosoares/Coffee-Future-Price-Prediction/blob/2044135ec7c5bcfb6f3

#read file XLSX format with decisions
ICFFUT <- read_xlsx("CoffeDatabase.xlsx", sheet = "ICFFUT")
#read file XLSX format with ICF Prices
ICFFUT_XLSX <- read_xlsx("CoffeDatabase.xlsx", sheet = "ICFFUT_XTS")
#Convert XLXS format to Data Frame
DFICFFUT_XTS <- as.data.frame(ICFFUT_XLSX)
#Convert Data Frame to XTS
ICFFUT_XTS <- xts(DFICFFUT_XTS[-1], order.by = as.Date(DFICFFUT_XTS$Date))
```

Extensions: moving average 5 days; moving average 22 days; Bollinger Bands; Rate of Change Oscillator; Relative Strength Index; Stochastic Momentum Index; MACD Oscillator.

```
##Technical Indicators - Price-Based
#moving average 5 days
Moving_Average_5 <- SMA(ICFFUT_XTS$ICFFUT.Close, n=5)

#moving average 22 days
Moving_Average_22 <- SMA(ICFFUT_XTS$ICFFUT.Close, n=22)

#Bollinger Bands
Bollinger_Bands <- BBands(ICFFUT_XTS$ICFFUT.Close)

#Rate of Change Oscillator
Rate_Change_Oscillator <- ROC(ICFFUT_XTS$ICFFUT.Close, n=10)

#Relative Strength Index
Relative_Strength_Index <- RSI(ICFFUT_XTS$ICFFUT.Close)

#Stochastic Oscillator / Stochastic Momentum Index
Stochastic <- stoch(ICFFUT_XTS$ICFFUT.Close)

#MACD Oscillator
MACD <- MACD(ICFFUT_XTS$ICFFUT.Close)
```

The raw data and the extensions are merged to create the complete data frame.

```
#merge XTS
Full_ICFFUT <- merge(ICFFUT_XTS,
                     Moving_Average_5,
                     Moving_Average_22,
                     Bollinger_Bands,
                     Rate_Change_Oscillator,
                     Relative_Strength_Index,
                     Stochastic,
                     MACD)

#create data frame ICFFUT - 4/5 Arabica Coffee Futures
```

```
DFICFFUT <- data.frame(Date=index(Full_ICFFUT), coredata(Full_ICFFUT))

#Merge Data frame ICFFUT - 4/5 Arabica Coffee Futures and ICFFUT
CDFICFFUT <- left_join(DFICFFUT,
                       ICFFUT |> select(Date, Decision, OpenInterest)) |>
  filter(Date > "2003-02-17") |> #excluding NA
  mutate(Decision  = as.factor(Decision)) #Decision as a factor = Y
```

.

Data exploration and visualization

The exploration and visualization of the data provide insightful information about the users, the movies and ratings. The first 5 rows of the dataset that we use in this project is presented in table 1. The columns `movieId`, `userId` and `rating` are the variable of interest.

The variables, `movieId`, `userId` and `rating`, that is going to be used to build the model is presented as a matrix in the figure 1. The figure have the movies (`moveId`) in the x axis, the users in the y axis (`userId`), and the respective rating (`rating`). The matrix, extract from a sample of 50 users and 50 movies, provide some insights about the behavior of some users, the preference for some movies, and the sparsity of the matrix. The goal of this project is to fill the blank spaces with a rate.

To see a potential flaw in the data we make a slice of the top 5 most rated movies and most active users. The unique users that provided ratings, the unique movies that were rated and the unique rating provided by a unique user to a unique movie, are presented to illustrate the possible rating matrix $users \times movies$ $= 10677 \times 69878 = 746087406$ and the realized rating matrix 10000054, or 1.34% of points of the matrix is filled. The extremes values confirm that some users are much more actives than others, the most active user rated more than 50% of the total unique movies, and some movies have been rated for more than 1/3 of the total unique users.

The dataset distribution presented trough histograms provide some insights about the general proprieties of the data. As showed in the slice before some movies get rated more than others, and some user are more active than others.

Modeling approach

The recommendation system is better as its error has decreased, for this project the error is the typical error we make when predicting a movie rating $(\hat{y}_{u,i} - y_{u,i})$. The loss function used to evaluate the models is based on the residual mean squared error ($RMSE$) on a test set. The definition of $RMSE$ includes N , the number of user/movie combination, and the sum occurring over all these combination. In this case the Loss function is:

$$F1 = \frac{1}{\frac{\beta^2}{1+\beta^2}\frac{1}{\text{recall}} + \frac{1}{1+\beta^2}\frac{1}{\text{precision}}}$$

The model which $RMSE = 0$ is a perfect model prediction, or without errors. For this project the reported $RMSE < 0.86490$ is the goal. The $RMSE > 1$ means our error is larger than one star, which means a bad model.

Machine Learning

"If we have enough data, I know we can make predictions," Simons told a colleague.

Zuckerman, Gregory. The Man Who Solved the Market (p. 2). Penguin Publishing Group. Edição do Kindle.

The machine learning decisions are based on algorithms build with data, so for this project the dataset XXXXX are going to be used to train and test the model. The train_set and test_set are build trough function `createDataPartition` as presented:

k-Nearest Neighbour Classification approach

The simplest model to recommend a movie to any user, is the model that predict the same rate μ for all movies regardless of user. In the naive approach the variation of the differences is random, and the independent error $\epsilon_{u,i}$ centered at 0. The model looks like:

$$Y_{u,i} = \mu + \varepsilon_{u,i}$$

The average of all ratings is the estimate that minimize the $RMSE$, and if we fill the blank cells in the matrix with the μ we obtain the $\varepsilon_{u,i}$.

Recursive Partitioning and Regression Trees approach

The values of b_i and b_u that minimize the full model and the code to calculate the $RMSE$ is presented below.Finally we achieve our goal, the $RMSE$ is lower than 0.86490. But can we do better?

Classification and Regression with Random Forest approach

A matrix Factorization is the last approach to solve recommendation system problem presented in thisThe model improved substantially and we achieve the best result so far, we have our choice to train and test the dataset.

Results

This section presents the modeling results and discusses the model performance of the `movielens` data frame. The code below provided by HarvardX and create the `edx` and `validation` datasets that will be used to train and test our final algorithm.a $RMSE$ of 0.78.

Conclusion

The recommendation system, developed in this project, provide suggestions for items that are most pertinent to a particular user. The goal of this project was achieved and the $RMSE$ of matrix factorization of 0.78 is lower than our target 0.86490. A larger dataset could provide a even better model since we have more data to train, and is one of the limitations of this project. The addition of more variables could be a good outlook for future works since can capture more structures and latent factors in the data.

Evaluation metrics

Avoiding False Positives: A Protocol

Model that makes sense economically: Renaissance Technologies Trading Strategies

1- Mean Reversion Strategy: Buy if the close price move a certain level below their recent trend line, Sell if above certain level above and be Neutral if the price is too close of trend line.

2- Stochastic Strategy: Non linear machine learning Kernel

3- The Kelly Criterion: scientific gambling method Assumption: future returns depend on decisions toked on close price. The relationship between the characteristics of market and performance is largely unknown and probably time-varying. This is why ML can be useful: to detect some hidden patterns beyond the documented asset pricing anomalies. Moreover, dynamic training allows to adapt to changing market conditions.

1. Research Motivation: ex ante/ex post economic foundation
a) Does the model have a solid economic foundation?

Coffee It is one of the most important commodities traded globally and one of the most popular beverages in the world. World production is spread over 55 developing countries and is conducted by about 26 million producers, cultivated in tropical and subtropical areas.

The development of local organizations or instruments to help managing price risk, such as organized futures, in developing countries, like Brazil, is a important step to minimize the risk and challenges of small producers, since the larger and more liquids futures markets are located in the United States and Europe.

The Futures price volatility is large once coffee market has low price elasticity of demand and supply in the short run. In addition, previous studies have pointed to the existence of asymmetric price transmission within the coffee chain. Because of the concentration of roasting and retail sectors, retail prices respond faster to price increases than to price decreases, while farm prices follow more closely supply and demand conditions.

b) Did the economic foundation or hypothesis exist before the research was conducted?
Price changes are commonly caused by the arrival of new information from ICO crop reports, as presented in work The Reaction of Coffee Futures Price Volatility to Crop Reports (2017)

2. Multiple Testing and Statistical Methods
a) Did the researcher keep track of all models and variables that were tried (both successful and unsuccessful) and are the researchers aware of the multiple-testing issue?

b) Is there a full accounting of all possible interaction variables if interaction variables are used?

c) Did the researchers investigate all variables set out in the research agenda or did they cut the research as soon as they found a good model?

3. Data and Sample Choice
a) Do the data chosen for examination make sense? And, if other data are available, does it make sense to exclude these data?
b) Did the researchers take steps to ensure the integrity of the data?
c) Do the data transformations, such as scaling, make sense? Were they selected in advance? And are the results robust to minor changes in these transformations?
d) If outliers are excluded, are the exclusion rules reasonable?
e) If the data are winsorized, was there a good reason to do it? Was the winsorization rule chosen before the research was started? Was only one winsorization rule tried (as opposed to many)?

4. Cross-Validation
a) Are the researchers aware that true out-of-sample tests are only possible in live trading?
b) Are steps in place to eliminate the risk of out-of-sample “iterations” (i.e., an in-sample model that is later modified to fit out-of-sample data)?
c) Is the out-of-sample analysis representative of live trading? For example, are trading costs and data revisions taken into account?

5. Model Dynamics
a) Is the model resilient to structural change and have the researchers taken steps to minimize the overfitting of the model dynamics?
b) Does the analysis take into account the risk/likelihood of overcrowding in live trading?
c) Do researchers take steps to minimize the tweaking of a live model?

6. Complexity
a) Does the model avoid the curse of dimensionality?
b) Have the researchers taken steps to produce the simplest practicable model specification?
c) Is an attempt made to interpret the predictions of the machine learning model rather than using it as a black box?

7. Research Culture
a) Does the research culture reward quality of the science rather than finding the winning strategy?
b) Do the researchers and management understand that most tests will fail?
c) Are expectations clear (that researchers should seek the truth not just something that works) when research is delegated?

Category #1: Research Motivation

Category #2: Multiple Testing and Statistical Methods Keep track of: what is tried, combinations of variables, Beware the parallel universe problem.

Category #3: Sample Choice and Data Define the test sample ex ante. Ensure data quality Document choices in data transformations Do not arbitrarily exclude outliers. Select Winsorization level before constructing the model.

Category #4: Cross-Validation Acknowledge out of sample is not really out of sample Understand iterated out of sample is not out of sample. Do not ignore trading costs and fees.

Category #5: Model Dynamics Be aware of structural changes. Acknowledge the Heisenberg Uncertainty Principle and overcrowding. Refrain from tweaking the model.

Category #6: Model Complexity Beware the curse of dimensionality. Pursue simplicity and regularization. Seek interpretable machine learning.

Category #7: Research Culture Establish a research culture that rewards quality. Be careful with delegated research

References

Coqueret, Guillaume, and Tony Guida. 2020. *Machine learning for factor investing: R version.* Chapman; Hall/CRC.

Arnott, Robert D., Campbell R. Harvey, and Harry Markowitz. 2018. "A Backtesting Protocol in the Era of Machine Learning." *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.3275654.

Coqueret, Guillaume, and Tony Guida. 2020. *Machine Learning for Factor Investing.* Chapman; Hall/CRC. https://doi.org/10.1201/9781003034858.