SPATIAL

Soybean Price forecAsting based on saTellite-derIved services and Artificial inteLligence

Executive Summary Report

Prepared By

Hypertech

Prime contractor

IAASARS

Subcontractor

List of Figures

List of Tables

List of Acronyms and Abbreviations

ARIMA	Autoregressive integrated moving average
CDL	Cropland Data Layer
CNN	Convolutional Neural Network
CNY	Chinese yuan renminbi
DL	Deep Learning
EO	Earth Observation
ML	Machine Learning
NASS	National Agricultural Statistics Service
SP500	Standard and Poor's 500
SSE	Shanghai Stock Exchange
PoC	Proof of Concept Prototype
TF	TensorFlow
US	United States
USD	United States dollar
USDA	United States Department of Agriculture

Table of Contents

1. Introduction

1.1 Scope of the Document

The present document summarizes the key results and findings of the project "SPATIAL - Soybean Price forecAsting based on saTellite-derIved services and Artificial inteLligence", providing a brief overview of the whole program, major findings, conclusions and further study areas.

1.2 Background and aim of the study

The present executive summary report is delivered as part of the contract closure documentation for the SPATIAL project, funded by ESA under the ESA AO/1-10468/20/I-FvO FUTURE EO-1 EO SCIENCE FOR SOCIETY PERMANENTLY OPEN CALL FOR PROPOSALS.

SPATIAL's starting point was that the great majority of publicly available models for forecasting commodity prices is based on time series forecasting utilizing the commodity prices as the singular data input. Such forecast models do not provide adequate means for including data representing external factors that may have a significant impact on the time series, such as weather, crop status and expected production volume, other commodities prices or exchange rates etc. Univariate forecasting techniques, such as Holt-Winters and ARIMA, are widely adopted. However, univariate forecasting techniques, by definition, do not take into account multiple data sources. All these econometrics-based techniques that are still widely used suffer from poor forecasting performance due to inherent shortcomings such as bias introduced through omitting variables, not being able to cope with high-dimensional datasets with complex relations, missing out on non-linear interactions, false assumptions on feature importance, and not weighting errors by confidence. It has been shown in late years that such shortcomings are overcome when using ML-based methods and techniques to come up with better forecasting models. SPATIAL advocates that better forecasting models need to also assess the external factors related to crop status and expected volume. As such, SPATIAL considers the integration and assessment of a plethora of open and free Earth Observation (EO) datasets to forecast soyabeans yield/production that serves as an additional feature to the Soybeans Futures Contract prices forecasting ML-based models developed within the project. ERA5 Reanalysis Copernicus Climate Change Services (C3S), Landsat 7 and 8 data in the visible, near infrared and shortwave infrared parts of the E/M spectrum and Sentinel-2 data that offers additional bands in the vegetation Red Edge have been considered along with Sentinel-1 Synthetic Aperture Radar data which where necessary for also studying in-season crop mapping techniques.

Taking advantage of the existing market gap, SPATIAL- in its full-blown implementation as a service- has the ambition to disrupt the market of soybean commodities purchasing in Europe by providing small and medium sized European companies buying soybean commodities a powerful and affordable forecasting and risk management service to support them in their purchasing decisions and protect them from price volatility related losses. To this day, there is no competitive FinTech solution for providing affordable high-accuracy price forecasting for soy commodities.

2. Key Findings

3.1 ML models for In-season crop mapping

Crop rotation practices together with the target to scale up globally have dictated the necessity to develop in-season crop mapping algorithms. The objective of this module is to detect the soybeans crops during the running season, using several Machine Learning and Deep Learning categorical classification algorithms. The In-season crop mapping utilises Sentinel-1 and -2, active and passive remote sensing datasets, respectively, to detect the soybeans crops during the running season. It also uses reference data for training and validation from USDA NASS CDL and the Tiger DB for county boundaries. As season progresses and more data are collected the accuracy of the result is increased. Intensive testing showed that adequate predictions were achieved in late summer months. The model developed and trained can be transferred to other areas in the world that lack detailed reference crop type datasets ("transfer learning"). The ML algorithm that produced the best results was an ensemble model of Gradient Boosting, Random Forest and Logistic Regression that takes hard decisions depending on the values produced by the 3 classifiers. The results of overall accuracy reached values ranging from 82.5% to 90.3%.

3.2 ML models for Soybean yields/production prediction

CNNs that have prevailed and are popular in deep learning is the implementation of TensorFlow (TF) from Google and PyTorch or simply Torch by Facebook. Both packages were tested in SPATIAL and produced similar results using identical CNNs. In all our analysis the results of TF are given. To produce efficient results (with respect to performance and accuracy) using the optimization process (aided of course by backpropagation) the original dataset has to be introduced in the training procedure in batches. This technique leverages accuracy and speed. The size of batch fed to the neural net fitting method needs to be tested to produce robust outcomes. The number of epochs (iterations) used is another important parameter to evaluate using relevant experiments. Numerous experiments were carried out to reveal the importance of the features used, the time-periods taken, and the operational and timely availability of satellite images. Integration of the data from August played a decisive role: firstly because the soybeans mask can be updated with confidence at that time and secondly because the forecast itself is more accurate and relies on the updated mask. An interesting year was 2019, with a record cool and wet spring, that caused both models to perform worse. Conversely, the inclusion of this unusual year in the training set, increased the performance of the models in 2020. The average result of all the experiments for different periods is 4.16 billion bushels when the reference value is 4.135. This is a positive deviation of 0.5%, which increases when subsets of features are used.

3.3 ML models for Soybean Futures prices forecasting

The objective of this module is to detect the predictability in Soybean Future Contracts price. A series of features to feed the machine learning models was utilized, including other relative commodities as well as other asset classes, such as currency pairs, stock indices and production related features. An important contribution to the features family was the yield and the production forecasts as it is described in section 3.2. The Machine Learning models that have

been developed to deal with the binary classification problem of predicting the direction of Soybean Future Contracts price include a Logistic Regression model, a Support Vector Machine model, a Random Forest model, an Extreme Gradient Boost model, an Artificial Neural Network model and an Ensemble Learning model. These models provided decent-to-very-good results of prediction accuracy in most of the requested time frames. The Normal Distribution Theory and the Efficient Market Hypothesis were used to extend the prediction from a single quantified direction (price up or down) to a table of price ranges with respective probabilities. In the table below, one can see the best performing model for each timeframe.

Table 1: ML Models for Soybean Futures prices forecasting with the best performance for each time horizon

Time horizon	*Best Model*	*Accuracy (%)*
1-week	Extreme Gradient Boost	55.73
2-weeks	Extreme Gradient Boost	63.36
1-month	Logistic Regression	63.36
2-months	Extreme Gradient Boost	59.16
3-months	Support Vector Machine	60.31
6-months	Extreme Gradient Boost	44.65

3.4 Explainable AI

Explainable Artificial Intelligence addresses the extremely important issue of correctly interpreting a prediction model's output and supporting understanding of the process being modeled. The state-of-the-art in explainable AI methodologies are the so-called Shapley values. SHAP values (SHapley Additive exPlanations) is a method based on cooperative game theory and used to increase transparency and interpretability of machine learning models. SHAP values break down a prediction to show the impact of each feature. SHAP values interpret the impact of having a certain value for a given feature in comparison to the prediction one would make if that feature took some baseline value. The evaluation of Shapley values for SPATIAL ML and DL models and datasets was based on the popular Python shap library for both ML as well as DL models.

For the crop yield forecasting Climate together with Sentinel-2 data are in the top six most important and impactful features for the success of the Soybean yields/production forecast. For the futures prices forecasting model the most important and impactful features comprise of SP500 Returns, SP500 Volume, Soybean Oil Returns, CNY/USD currency pair Returns, Crude Oil Returns, SSE Volume and Soybean Yield Prediction.

3.5 Proof of concept prototype

SPATIAL produced a Proof-of-Concept Prototype (PoC) which can be used to validate and confirm the predictability of the Soybeans Future contracts. The machine learning models developed have been executed offline for the period 1/1/2020 to 31/12/2020 and their results have been loaded into the PoC's Data Base. By using the PoC's Front-End web application, the user is able to ask for a prediction by providing a certain reference date within the above-mentioned period and prediction timeframe. The PoC then returns the prediction as a table

with price ranges and probabilities, the prediction result (successful or unsuccessful) and a series of other metadata, so the user can understand the quality and the success of the prediction, having in the same time, useful information regarding the economic conditions that take place at the time of the reference date. Finally, the PoC presents an analysis on how the prediction occurred, focusing on the contribution of each feature to support the Explainable AI paradigm. The Front-End of the PoC is a web application, developed with REACT framework. Selected views of the PoC User Interface are shown in Figure 1.

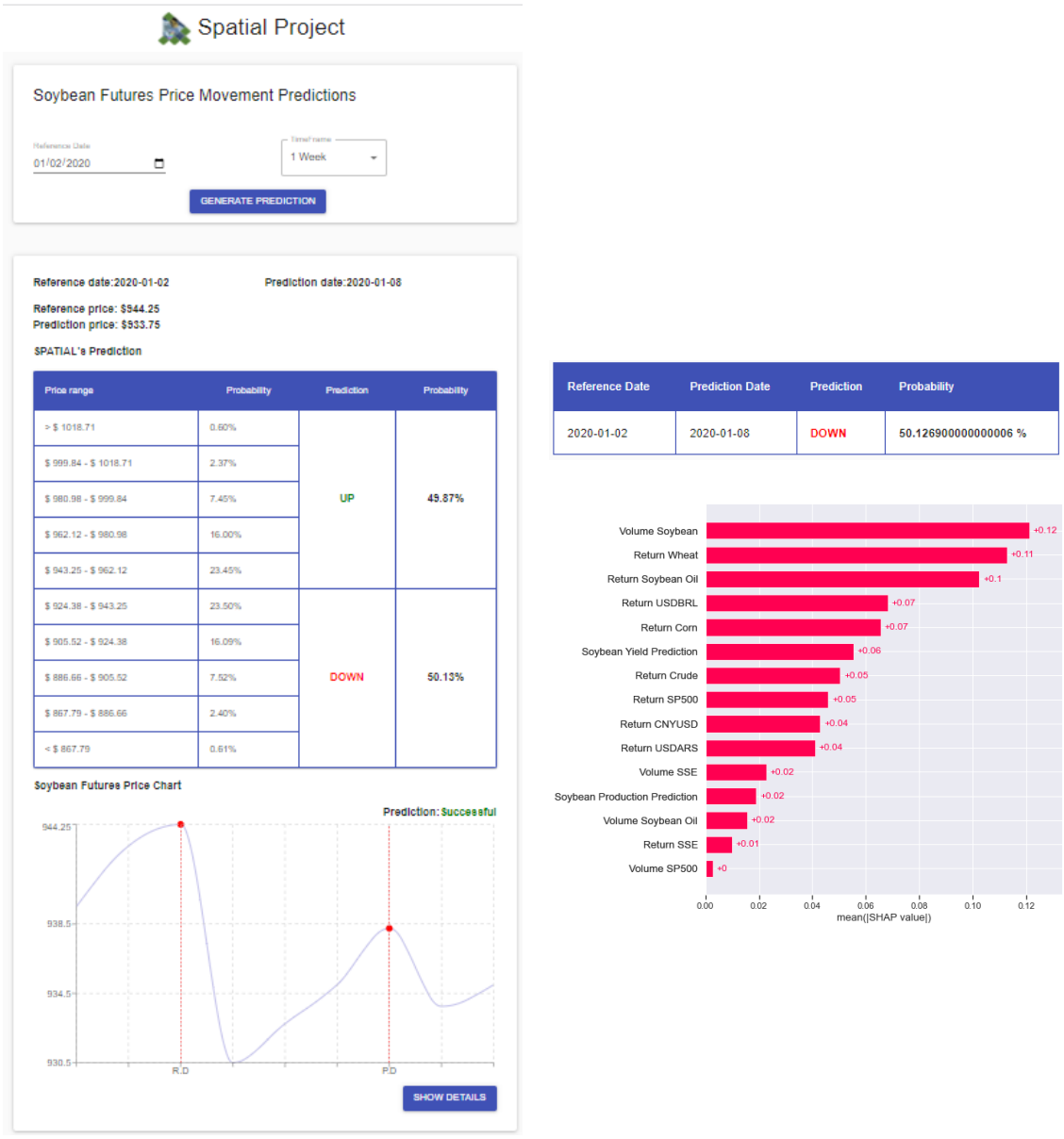

Figure 1: Selected views of the SPATIA Proof-of-Concept prototype

3. Conclusions

The problem of predicting financial time series is one of the most difficult problems in the Machine Learning community. The main reason for that is that these time series are affected very quickly, by a huge number of factors from millions of traders and investors all over the world that act as speculators in a zero-sum game. Although all the investors, either individual investors or institutional investors such as investment banks and hedge funds are trying to be profitable, not all of them act completely rationally because they have different ways of making decisions, including fundamentals, technical indicators and advanced algorithms. The difficulty of this venture can be shown due to the limited papers and bibliography from scientists that tried to address this problem.

As per the prediction of commodities' prices, things are even more complicated because there are other factors that affect the movement of the prices. All commodities besides the usual fundamentals depend on the supply, the demand and the currency markets because they are always denominated in a currency. The modern approach to this problem is to combine fundamentals with technical analysis in new techniques and algorithms such as Machine Learning Models. This is what has been done in the current project. SPATIAL developed several Machine Learning models trying to achieve classification predictability in the Soybeans Future Contracts prices by taking advantage of deep knowledge in technical analysis and the fundamentals. Since the arithmetic prediction – regression is unrealistic in this context, the focus has been on classification and more specifically binary classification, to substantiate that the Soybeans Future Contracts prices are predictable in terms of the price direction under certain timeframes. The key aim of the work performed was to combine the expertise in financial markets, gathering all the possible features (technical and fundamentals) that may affect prices with the unique EO-derived information regarding the yield and the production prediction.

Overall, the results of the machine learning models developed were very positive. The performance evaluation of the Proof of Concept prototype which integrated all the developed models revealed that, the short-term (1-week, 2-weeks, 1-month) and mid-term (2-months, 3-months) predictions generated significant results indicating the presence of predictable patterns in the Soybean Futures contract market. For the longer-term predictions (6-months), there doesn't seem to be any predictability. Nonetheless, this is quite expected given the very limited number of observations at the 6-months' time-horizon as well as the difficulty to predict financial time series in such a long time period.

Prediction accuracy above 57%-58% is considered as something beyond randomness and this is something that was achieved almost in all the requested time frames. Moreover, a methodology was developed that converts the direction prediction with its probability to a table that contains price ranges and it can be capitalized by traders and investors.

Since the predictability is validated with accuracy that sometimes exceeds 60%, as part of future work foreseen tasks are:

- Development of trading strategies that take advantage of the predictions generated by SPATIAL
- Application of risk management models to maximize the profit and to minimize the risk
- Extension to other commodities that can be predicted (e.g. Wheat, Maize), taking advantage of the EO-based yield & production predictions and a possible creation of a portfolio of assets.
- Commercial appropriation to help traders (individuals and companies) becoming more profitable.