Alternative data for ML-based asset performance forecasting

A Project Work, presented as part of the requirements for the Award of an International Master's Degree in Fintech, Finance and Digital Innovation from the MIP – Politecnico di Milano Graduate School of Business

José Guilherme Monteiro and Nicola Tombini

A Project carried out in collaboration with **Axyon AI** and under the supervision of:

Professor Daniele Marazzina

17/09/2020

Abstract

Title:	Alternative data for ML-based asset performance forecasting
Authors:	José Guilherme Monteiro, Nicola Tombini
Keywords:	Artificial Intelligence, Alternative Data, Deep learning, Finance, Fintech, Machine Learning

This project examines the possibility of inclusion of more Alternative Data sources into Axyon AI Machine Learning predictive models. Alternative Data is a type of data that has now, more than ever, been on the radars of Wall Street. This study investigates the potential of this Data, using backtesting to measure the impact on the performance of the forecasting. The results are based on two different datasets, one comprising Alternative Data previous selected, and the other with the usual standard data already used on the Axyon ML-platform. The findings reveal the use of alternative data to forecast asset-performance is significantly better than considering only the conventional data, specially after the beginning of the COVID-19 pandemic, on recent periods of high volatility. Moreover, further improvements are suggested even though implementation challenges are something also to take in future considerations and not covered on this research. By applying Alternative Data, the Axyon AI model was improved, which ultimately can have a positive impact on the returns of the investors' portfolios using Axyon technology.

Table of Contents

Glossary

AI	Artificial Intelligence
ANN	Artificial Neural Networks
DL	Deep Learning
GA	Generic Algorithm
GDP	Gross Domestic Product
LSTM	Long-Short Term Memory
ML	Machine Learning
NN	Neural Networks
US	The United States of America

List of figures

1. Introduction

1.1. Context and purpose of the study

Over the past few years, the use of ML-based techniques for financial forecasting has been significantly increasing. It is commonly said that nowadays Data drives the business, this statement is particularly true for companies which use Machine Learning on their core products, like Axyon AI. In ML often, the results obtained are proportionally connected with the quality of the data which feds the model. In specific, Alternative Data is a type of data that has been more and more on the radars of Wall Street and investment companies. Eagle Alpha founder and CEO, Emmett Kilduff to Wall Street Journal (2019), said the following on Alternative Data: "Using that gets you inside the boardroom of the company,".

The project aims to study if Axyon predictive models forecasting performance can benefit from the use of better or additional Alternative Data Sources.

1.2. Brief description of the company (Axyon) and its projects/products

This project work has been developed in collaboration with Axyon AI, a Modena-based company which is on a mission to bring asset management to the future with superior, advanced, accurate and consistent AI and deep learning predictive solutions. One of the company's solution is called "Axyon IRIS Forecasts Feeds" to help quantitative teams and portfolio managers by delivering AI-powered forecasts on the performance of target asset pools. Axyon IRIS is an AI engine and web application that uses proprietary AI and deep learning technology to deliver consistent and highly accurate forecasts on assets future performances. The idea behind this solution is to identify the probability of individual assets to outperform others over each prediction horizon.

2. Background information and Literature Review

2.1 Artificial intelligence

2.1.1. Concept and History

What is artificial intelligence? Whereas in academia there is no consensus of a broadly accepted definition of the term, Chrisley with Begeer (2000) suggests the intellectual strand begun in the seventeenth century with Descartes and Vaucanson, winding through other personalities such as Babbage, Boole, Whitehead Russell, Shannon, Turing, McCulloch and Pitts, Wiener and von Neumann (Chrisley with Begeer, 2000). Creatives and great thinkers in the past have always dreamed of conceiving an artificial object that could be able to think. Once computers were first invented, people wondered whether such machines might become intelligent, over a hundred years before one was built (Lovelace, 1842). The term artificial intelligence (AI) was defined by John McCarthy in 1956 (Russel and Norvig, 1995), some researchers define AI as the ability of machines to understand, act and learn like a human, in other words allowing the possibility of simulating human intelligence through the use of computational machines. While others, prefer to classify the AI power as the field which allows machines to act and think rationally 1956 (Russel and Norvig, 1995).

After so many years, still, no one entirely agrees on one single technical definition. However, it is reasonably hard to have a constructive debate about artificial intelligence (AI) without agreeing in standard general lines. A recent view of AI describes the concept for nowadays as "a thriving field with many practical applications and active research topics. Intelligent software is helpful to automate routine labour, understand speech or images, make diagnoses in medicine and support basic scientific research." (Goodfellow, Bengio and Courville, 2016).

2.1.2. AI in Finance

Several industries are focusing on artificial intelligence research; Finance is no exception. According to the World Economic Forum report, *The New Physics of Financial Services: Understanding how artificial intelligence is transforming the financial ecosystem* produced by Deloitte in 2018, "the future of financial services lies in its ability to fully benefit from technologies". AI as a suite of technologies, throughout its adaptive predictive and learning power, has the ability to dramatically enhance the human ability to: recognise patterns, anticipate future events, create good rules, make right decisions and communicate. The AI revolution is changing at a fast pace the traditional "recipe" to build a successful enterprise in financial services.

During the last few years, AI has been making significant progress towards the financial world. Especially through its sub-category Machine Learning (ML), enabled a revolution that is starting to gain a considerable dimension, and it is already disrupting the finance industry. In fact, some authors on the field like Lopez de Prado (2018) defend ML as transforming every aspect of our lives and that Financial ML is a distinct subject, related but separate from standard ML. It is also defended that ML will dominate Finance; science will overcome intuition and guessing; investing will not mean gambling.

2.2. Basic concepts and context around Deep Learning?

2.2.1. Machine learning overview

Computers are now able to tackle real-world problems and make subjective decisions. The challenges whose hard-coded systems face, indicate that AI systems must have the ability to acquire their knowledge, by writing patterns from raw data, this ability is known as machine learning (Goodfellow, Bengio and Courville, 2016).

Machine learning (ML) is a subset of AI, and ML algorithms are conventionally categorised as **unsupervised** or **supervised**, depending on the type of experience they are allowed to have during the learning process. The former is a learning approach where no labels are given, throughout a dataset with many features, then learn insightful characteristics of the structure of the dataset. In contrast, the later, supervised ML provides the algorithm with a training dataset with a label/target. There are also some ML algorithms which do not experience a fixed data set, it is the case of **reinforcement learning** algorithms, working on a reward loop system for training (Goodfellow, Bengio and Courville, 2016).

2.2.2. Artificial Neural Network models

Artificial Neural Networks (ANN) and its development rose from the attempt to simulate biological nervous systems by combining several simple computing elements, the neurons, into a too interconnected system, with the hope such phenomena as "intelligence" would surge from the result of self-organisation or learning (Sarle, 1994).

The structure of Neural Networks can be widely different; however, researchers usually identify five main types of ANNs from the most common to the least: Feedforward Networks, Recurrent Networks, Support Vector Machine, Modular Networks and Polynomial Networks (Gómez-Ramos and Francisco Venegas-Martínez, 2013).

2.2.3. Deep learning

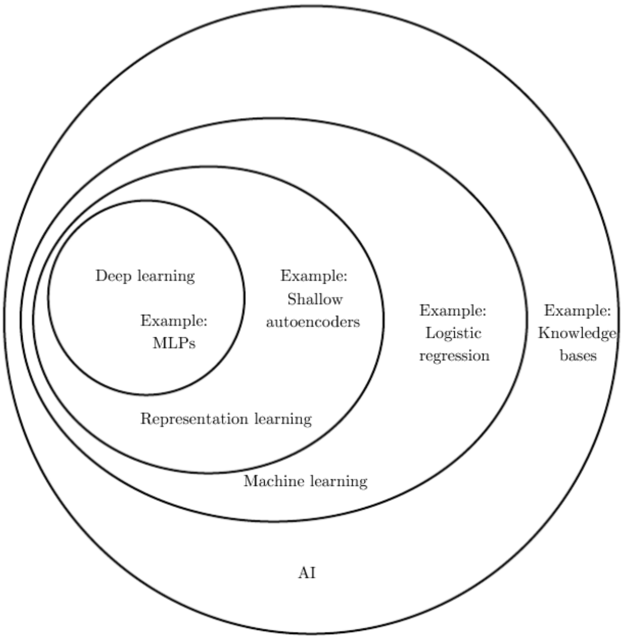


Figure 1. A Venn diagram of the AI world and some of its subsets, Machine Learning and Deep Learning. Each section of the Venn diagram includes an example of an AI technology. (Goodfellow, Bengio and Courville, 2016).

Deep learning (DL) is a particular kind of Machine Learning. In order to have a better understanding of where to locate DL on the AI world, it is possible to have this overview on Figure1 which provides a Venn diagram showing how deep learning is a kind of representation learning, which is, in turn, a kind of machine learning, which is used for many but not all approaches to AI. Deep Learning is described as a technique that enables computer systems to improve with experience and data, by introducing representations that are expressed in terms of other, simpler representations, building complex concepts out of simpler concepts (Goodfellow, Bengio and Courville, 2016). DL covers all three types of learning and has as its prominent example of the DL model in the **feed-forward deep network**, or **multi-layer perceptron** (MLP).

2.2.4. Multi-layer Perceptron (MLP)

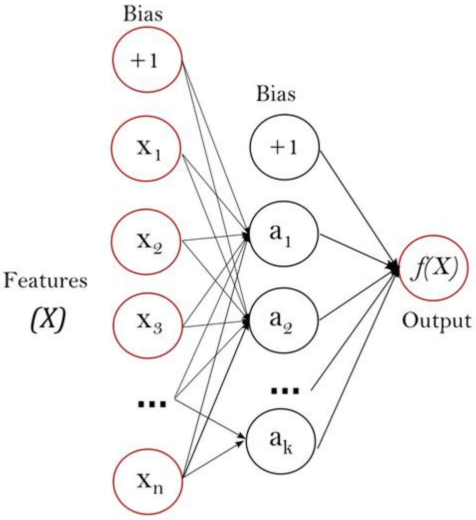


Figure 2. One hidden layer MLP (scikit-learn 0.23.2 documentation).

Multi-layer Perceptron (MLP) is a supervised learning algorithm, probably the most classic deep learning model. MLP is a feed-forward system due to the direction of the information flow. Information in feed-forward neural networks always flows through connections with neurons in the forward positions and never with the same or previous layers. In these networks, inversely from reinforcement learning, there is no feedback systems, when feed-forward neural networks are extended to include feedback connections, they are called **recurrent neural networks** (Goodfellow, Bengio and Courville, 2016). The MLP structure is different from logistic regression; it consists of an **input layer**, also called a **visible layer** because it contains the variables that can be observed, Then, between the input and output layer, one or multiple non-linear layers can be founded, denominated hidden-layers (Pedregosa, 2011). These layers are called "hidden" because their values are not given in the data, as an alternative, the model computes which concepts are useful for explaining the relationships in the observed data

(Goodfellow, Bengio and Courville, 2016). As described in Figure 2, the process can be briefly described by the following: "The leftmost layer, known as the input layer, consists of a set of neurons $\{x_i|x_1, x_2, \ldots, x_m\}$ representing the input features. Each neuron in the hidden layer transforms the values from the previous layer with a weighted linear summation $w_1x_1 + w_2x_2 + \ldots + w_mx_m$, followed by a non-linear activation function $g(\cdot): R \rightarrow R$ - like the hyperbolic tan function. The output layer receives the values from the last hidden layer and transforms them into output values" (Pedregosa, 2011).

2.3. Alternative Data

2.3.1. The use of the term in Finance

Although there is no academic definition of alternative data, the term, in the finance context, refers to the data used in the investment decision process, with profit maximisation purpose, that is outside of the conventional financial data sources (e.g. SEC filings, earnings releases, financial statements, industry benchmarks etc.). Kolanovic and Krishnamachari (2017) divided financial alternative data in three groups, data produced by individuals (social media, news, web searches, etc.), business processes (transactions, corporate data, government agencies, etc.), and sensors (satellites, geolocation, weather, CCTV, etc.). Simon Constable on an article for Wall Street Journal (2019) offers an (also) alternative definition of the term: "Alternative data in Finance refers to proxy metrics or information originating from unofficial or noncompany sources that individuals can use to gain insight into an investment. Such information can be the difference between making a profitable and unprofitable bet."

2.3.2. Problematic aspects and adoption

Lopez de Prado (2018) states that what most defines alternative Data is that it is primary information that has not reached other sources yet. He adds two problematic aspects for this type of data; it is costly and has privacy concerns associated as well. The high elevated costs associated with this particular kind of data seems that is not slowing down financial companies from purchasing and chasing alternative data. According to a data provider (Dataiku, 2020), buy-side spending on alternative data has increased from $232 million in 2016 to more than $650 million in 2018, and spending is estimated to exceed $1.7 billion by the end of 2020.

2.3.3. Alternative Data in the age of coronavirus

The unprecedented times the whole world is living in, presented the world with a global turmoil, caused by the pandemic, which is having a massive impact on the world's global economy.

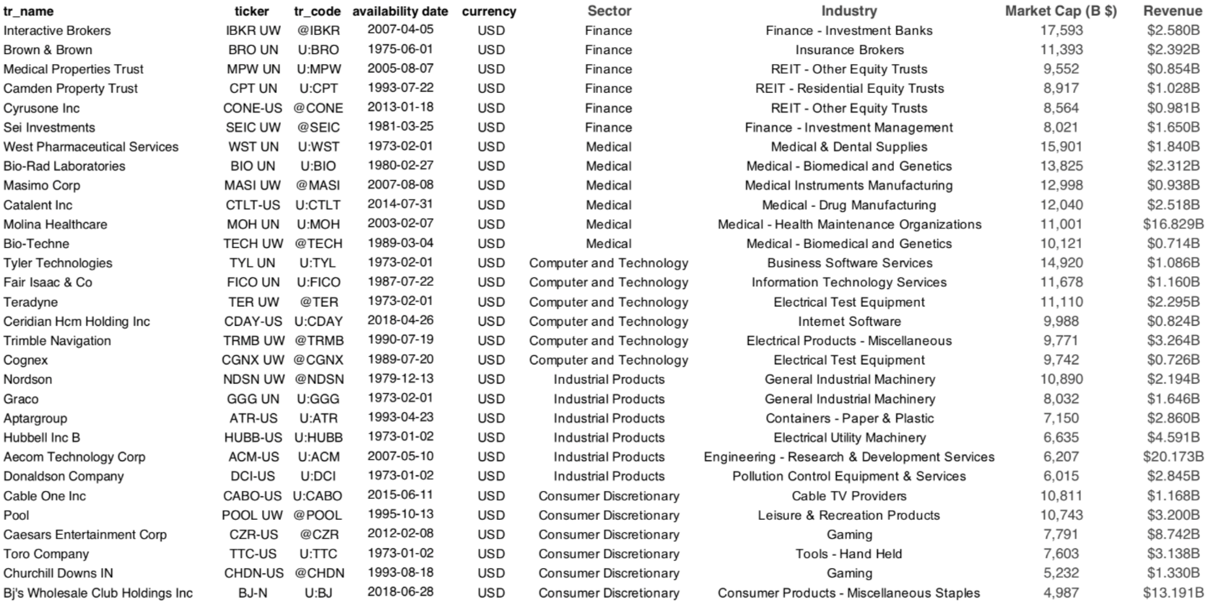
With the S&P500 hitting record highs while Fed officials said the coronavirus pandemic will continue to "weigh heavily" on the U.S. economy, seems to happen a recent disconnection between the stock markets and the so-called "real economy". Perhaps due to the phenomenon previously mentioned, the truth is that investors and analysts are turning to a variety of unconventional data to forecast where the market is heading.

Noah Smith, in a Bloomberg Opinion article (2020), suggests two major data problems that the age of coronavirus has brought to investors, analysts, journalists and researchers: "First, traditional government numbers such as unemployment and gross domestic product growth only come out once a month or once a quarter, making it hard to spot fast-changing trends as they occur. Second, the pandemic produces a lot of unusual economic effects that make traditional numbers harder to interpret." In response to these problems, many people are looking for a solution near alternative data sources.

On an article entitled by "Wall Street's New Metrics" from Wall Street Journal (Otani and Wursthorn, 2020), it is enhanced the increase in demand for alternative metrics. "Traders and analysts track unconventional indicators as they look for market's trajectory", they said. Some examples, given in the article, for the most commonly cited new metrics are restaurant reservations, travel through airports, retail-store traffic, searches for directions (e.g. mobility data from Apple Inc.) or the unfortunately famous, number of new COVID-19 cases.

3. Data

This chapter describes the data used in the project. To our research, a sample of 30 US Mid Cap Stocks from the Russell 2000 index was chosen, as it is possible to observe on the Figure 3. This study uses two different datasets: a so-called "original dataset" which contains the usual features already used in Axyon machine learning – platform, and a second dataset consisting of the previous dataset but with the additional implementation of Alternative Data. This chapter is, therefore, divided in two main parts, each of them representing the respective dataset with a stronger emphasis on the Alternative Data one, describing its content and the selection process.



tr_name	ticker	tr_code	availability date	currency	Sector	Industry	Market Cap (B $)	Revenue
Interactive Brokers	IBKR UW	@IBKR	2007-04-05	USD	Finance	Finance - Investment Banks	17,593	$2.580B
Brown & Brown	BRO UN	U:BRO	1975-06-01	USD	Finance	Insurance Brokers	11,393	$2.392B
Medical Properties Trust	MPW UN	U:MPW	2005-08-07	USD	Finance	REIT - Other Equity Trusts	9,552	$0.854B
Camden Property Trust	CPT UN	U:CPT	1993-07-22	USD	Finance	REIT - Residential Equity Trusts	8,917	$1.028B
Cyrusone Inc	CONE-US	@CONE	2013-01-18	USD	Finance	REIT - Other Equity Trusts	8,564	$0.981B
Sei Investments	SEIC UW	@SEIC	1981-03-25	USD	Finance	Finance - Investment Management	8,021	$1.650B
West Pharmaceutical Services	WST UN	U:WST	1973-02-01	USD	Medical	Medical & Dental Supplies	15,901	$1.840B
Bio-Rad Laboratories	BIO UN	U:BIO	1980-02-27	USD	Medical	Medical - Biomedical and Genetics	13,825	$2.312B
Masimo Corp	MASI UW	@MASI	2007-08-08	USD	Medical	Medical Instruments Manufacturing	12,998	$0.938B
Catalent Inc	CTLT-US	U:CTLT	2014-07-31	USD	Medical	Medical - Drug Manufacturing	12,040	$2.518B
Molina Healthcare	MOH UN	U:MOH	2003-02-07	USD	Medical	Medical - Health Maintenance Organizations	11,001	$16.829B
Bio-Techne	TECH UW	@TECH	1989-03-04	USD	Medical	Medical - Biomedical and Genetics	10,121	$0.714B
Tyler Technologies	TYL UN	U:TYL	1973-02-01	USD	Computer and Technology	Business Software Services	14,920	$1.086B
Fair Isaac & Co	FICO UN	U:FICO	1987-07-22	USD	Computer and Technology	Information Technology Services	11,678	$1.160B
Teradyne	TER UW	@TER	1973-02-01	USD	Computer and Technology	Electrical Test Equipment	11,110	$2.295B
Ceridian Hcm Holding Inc	CDAY-US	U:CDAY	2018-04-26	USD	Computer and Technology	Internet Software	9,988	$0.824B
Trimble Navigation	TRMB UW	@TRMB	1990-07-19	USD	Computer and Technology	Electrical Products - Miscellaneous	9,771	$3.264B
Cognex	CGNX UW	@CGNX	1989-07-20	USD	Computer and Technology	Electrical Test Equipment	9,742	$0.726B
Nordson	NDSN UW	@NDSN	1979-12-13	USD	Industrial Products	General Industrial Machinery	10,890	$2.194B
Graco	GGG UN	U:GGG	1973-02-01	USD	Industrial Products	General Industrial Machinery	8,032	$1.646B
Aptargroup	ATR-US	U:ATR	1993-04-23	USD	Industrial Products	Containers - Paper & Plastic	7,150	$2.860B
Hubbell Inc B	HUBB-US	U:HUBB	1973-01-02	USD	Industrial Products	Electrical Utility Machinery	6,635	$4.591B
Aecom Technology Corp	ACM-US	U:ACM	2007-05-10	USD	Industrial Products	Engineering - Research & Development Services	6,207	$20.173B
Donaldson Company	DCI-US	U:DCI	1973-01-02	USD	Industrial Products	Pollution Control Equipment & Services	6,015	$2.845B
Cable One Inc	CABO-US	U:CABO	2015-06-11	USD	Consumer Discretionary	Cable TV Providers	10,811	$1.168B
Pool	POOL UW	@POOL	1995-10-13	USD	Consumer Discretionary	Leisure & Recreation Products	10,743	$3.200B
Caesars Entertainment Corp	CZR-US	@CZR	2012-02-08	USD	Consumer Discretionary	Gaming	7,791	$8.742B
Toro Company	TTC-US	U:TTC	1973-01-02	USD	Consumer Discretionary	Tools - Hand Held	7,603	$3.138B
Churchill Downs IN	CHDN-US	@CHDN	1993-08-18	USD	Consumer Discretionary	Gaming	5,232	$1.330B
Bj's Wholesale Club Holdings Inc	BJ-N	U:BJ	2018-06-28	USD	Consumer Discretionary	Consumer Products - Miscellaneous Staples	4,987	$13.191B

Figure 3. Target Assets – 30 US Mid Cap Stocks (Russell 2000 Index)

3.1. Dataset 1 description

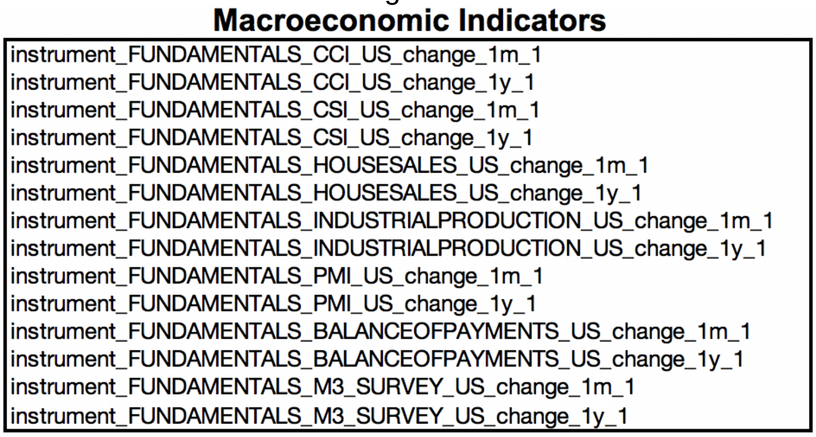
3.1.1. Standard dataset

The original dataset includes 65 features which are already stored inside Axyon database and downloaded from Refinitiv. They must be updated regularly to be used to make live predictions. Considering pair trading will be performed, the data set will have the double of the features and so 130 columns.

In particular, those features represent for each asset some input data (market data, fundamentals) and there are selected some features like:

- Technical indicators (e.g. rate of change, moving average);
- Categorical features (e.g. sector);
- Context features. (e.g. macroeconomic indicators)

On Figure 4 it is possible to observe a table showing the features representing the macroeconomic indicators included in the original dataset.



Macroeconomic Indicators
instrument_FUNDAMENTALS_CCI_US_change_1m_1
instrument_FUNDAMENTALS_CCI_US_change_1y_1
instrument_FUNDAMENTALS_CSI_US_change_1m_1
instrument_FUNDAMENTALS_CSI_US_change_1y_1
instrument_FUNDAMENTALS_HOUSESALES_US_change_1m_1
instrument_FUNDAMENTALS_HOUSESALES_US_change_1y_1
instrument_FUNDAMENTALS_INDUSTRIALPRODUCTION_US_change_1m_1
instrument_FUNDAMENTALS_INDUSTRIALPRODUCTION_US_change_1y_1
instrument_FUNDAMENTALS_PMI_US_change_1m_1
instrument_FUNDAMENTALS_PMI_US_change_1y_1
instrument_FUNDAMENTALS_BALANCEOFPAYMENTS_US_change_1m_1
instrument_FUNDAMENTALS_BALANCEOFPAYMENTS_US_change_1y_1
instrument_FUNDAMENTALS_M3_SURVEY_US_change_1m_1
instrument_FUNDAMENTALS_M3_SURVEY_US_change_1y_1

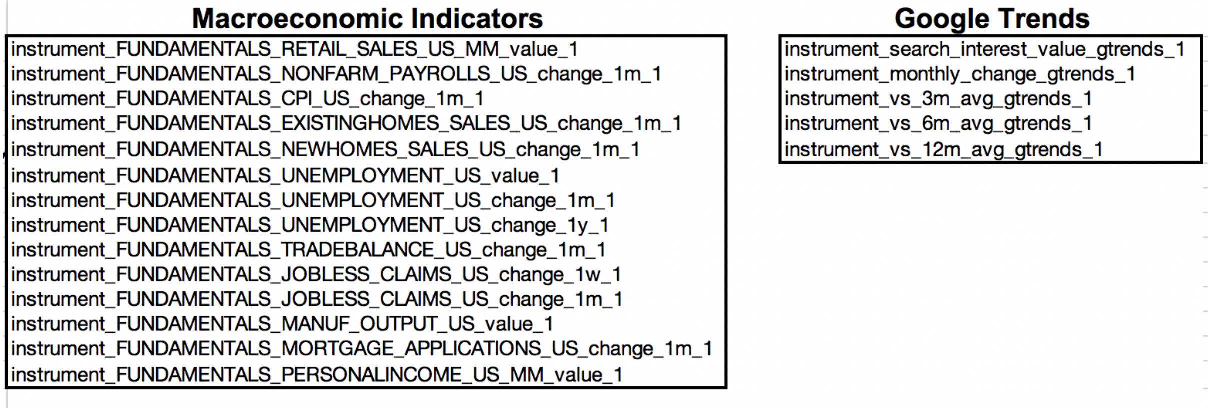
Figure 4. Macroeconomic Indicators - Standard Dataset

3.2. Dataset 2 description

3.2.1. Alternative Data dataset

The Alternative dataset shows 84 features, including traditional data, the same features as the original dataset, and in addition, alternative data, which is the primary Data of our study. The selection included some requirements, the most important criteria was the frequency of the data, which had to be daily or at least weekly and the necessity of having historical data for a period of at least 10 years. Concerning the alternative data, the focus is on an asset-specific data which is well represented by google search trend (of a specific asset). Also Global features were selected, which are mainly macro indicators, like "US Retail Sales" or "US unemployment rate", downloaded from Eikon Thomson Reuters, with an available time window from January-2010 until the end of August 2020 and a target horizon of 20 days. Considering a pair trading strategy will be performed, which will be described later, the dataset takes double features which corresponds to 168 columns. Further in this report will be explained in more depth the alternative data features selected for this dataset.

On Figure5, two tables are showing the features representing the alternative data which have been added to the original dataset to build the alternative dataset. It can be seen that they are 19, which is precisely the difference between the number of features in the alternative dataset (84I and in the original dataset (65).



Macroeconomic Indicators
instrument_FUNDAMENTALS_RETAIL_SALES_US_MM_value_1
instrument_FUNDAMENTALS_NONFARM_PAYROLLS_US_change_1m_1
instrument_FUNDAMENTALS_CPI_US_change_1m_1
instrument_FUNDAMENTALS_EXISTINGHOMES_SALES_US_change_1m_1
instrument_FUNDAMENTALS_NEWHOMES_SALES_US_change_1m_1
instrument_FUNDAMENTALS_UNEMPLOYMENT_US_value_1
instrument_FUNDAMENTALS_UNEMPLOYMENT_US_change_1m_1
instrument_FUNDAMENTALS_UNEMPLOYMENT_US_change_1y_1
instrument_FUNDAMENTALS_TRADEBALANCE_US_change_1m_1
instrument_FUNDAMENTALS_JOBLESS_CLAIMS_US_change_1w_1
instrument_FUNDAMENTALS_JOBLESS_CLAIMS_US_change_1m_1
instrument_FUNDAMENTALS_MANUF_OUTPUT_US_value_1
instrument_FUNDAMENTALS_MORTGAGE_APPLICATIONS_US_change_1m_1
instrument_FUNDAMENTALS_PERSONALINCOME_US_MM_value_1

Google Trends
instrument_search_interest_value_gtrends_1
instrument_monthly_change_gtrends_1
instrument_vs_3m_avg_gtrends_1
instrument_vs_6m_avg_gtrends_1
instrument_vs_12m_avg_gtrends_1

Figure 5. Alternative Features - Alternative Data Dataset

3.2.2. Constraints

Many problems were faced while chasing alternative data. The first obstacle was that many alternative data was paid and too expensive, and once data providers were contacted, in particular Quandl and Hanweck, getting free trials was not possible. Then another critical limitation was that looking for data with long term horizon, some data may be missing, and one option was either discard data, generally for asset-specific features, or fill it in somehow, for global features. In this framework, it was realised that using alternative data could face a big issue and so that the resulting dataset may be relatively small due to the time-frame availability of features considered. An excellent example of this is given by data related to COVID-19. As it is known, there was no data for COVID-19 before 2019. Indeed, if it was to include this feature, the dataset can be generated only starting from 2019, as it is not possible to train on empty data. In this case, COVID-19 data failed the requirements needed to perform our research.

3.3. Alternative Data selection: Google Trends, Macroeconomic Indicators

3.3.1. What is Google Trends?

"Google Trends is a search trend feature that shows how frequently a given search term is entered into Google's search engine relative to the site's total search volume over a given period of time. Google Trends can be used for comparative keyword research and to discover event-triggered spikes in keyword search volume."

In Finance, Google Trends can be used in investment strategy building an investor sentiment index which is able to measure bullish vs bearish sentiment. Indeed, analysing what people search on google, it is possible to understand if investors are becoming worried about the stock market and want to sell stocks ("how to short sell") or if they are confident of future good performance of the stock market ("how to invest").

In this project, it will be analysed how many times a stock belonging to our asset pool is searched on google in order to use those data to predict a possible future investment in a particular stock. During this analysis, there was a problem in terms of time-frame availability. The goal was to have google trends data of a specific asset updated daily for a time horizon of 10 years. Unfortunately, google allows importing daily data only up to 5 years. Indeed, for ten years of historical Data, it gives monthly data. At this point, monthly search trends data was added to our dataset. Attached to the Appendix, it is possible to find the code used on the Google Trends data extraction process.

3.3.2. Macro-Economic Indicators with Eikon data source

Concerning the global features that were added to the dataset as alternative data, helping asset performance forecasting to be more accurate, were considered 11 macro-economic indicators available in the Eikon data source. First of all, it is essential to note that Eikon carries around 1800 real-time economic indicators from around the globe. There are economic indicators where the releases in real-time over our real-time network are tracked and provide additional value-added information such as Reuters Polls as well as other useful information such as the significance of the economic indicator. Additionally, Eikon carries over 400,000 non-realtime economic indicators globally. In the near future, the number of non-realtime indicators will more than double to 1 million time series. This is a great deal of economic time series for an analyst to work with to generate insight and models potentially.

In the alternative dataset, were added 11 US macro-economic indicators available in Eikon. Below it is a brief description of the picked features:

- "US Retail Sales MM" is an important factor which may affect dollar quotes. Indeed, if it goes down, it reveals that consumers have reduced their spending level, leading to a decline in economic activity. In particular, it shows changes in the volume of US retail sales in the given month compared to the previous month. Two types of retail companies are taking into account in this monthly indicator: stores with fixed points of sale and without them (using paper and electronic catalogs, mobile stands, home-based sales, vending machines, etc.). (U.S. Census Bureau, 2020)
- "Non-Farm Payrolls" measures the change in the number of U.S workers during the previous month (excluding general government employees, private household employees, employees of nonprofit organisations that assist individuals, and farm employees). Job creation is the primary indicator of consumer spending, which accounts for the majority of economic activity. This indicator is analysed closely because of its importance in identifying trends related to the rate of economic growth and inflation. (Balasubramaniam, K., 2020)
- "Consumer Price Index (CPI)" is an indicator that measures the changes in the prices of a set of consumer goods and services such as transportation, food, and medical care. In particular, it looks at the weighted average of price changes for each item. These changes are often used as an essential factor to evaluate the cost of living. The CPI is one of the most important indexes to assess periods of inflation or deflation in the economy. (Chen, J., 2020)
- "Existing Home Sales" is an important indicator tracking regional transaction data in the U. S's existing stock of single-family homes, condos, and co-ops. It is a lagging indicator since people often make housing choices in response to a change in interest rates. (Kagan, J., 2020)
- "New Home Sales" is an economic indicator that measures sales of newly built homes. It is a factor which is tracked by investors to evaluate real estate market demand and monitor mortgage rates. It has to be taken into account that other factors, like household income, unemployment, and interest rates typically influence new residential sales. (Fernando, J., 2020)
- "Unemployment rate", is one of the main macro-economic indicators, assessed by a state to track its economy, which might be different concerning different parts of U.S. As the name suggests, it's computed by the ratio between the number of unemployed individuals in a state and the total labour force. (Williams, W., 2020)
- "Initial Jobless Claims" is a weekly indicator reported by U.S Department of Labor which tracks the number of new jobless claims filed by individuals for the first time during the

past week in order to receive unemployment benefits. An high value in the initial claims reveals a weakening economy and the beginning of a recession. (Kagan, J., 2020)

- "International trade" is a global indicator that measures the exchange of goods and services between countries. It's essential for the global economy it shows the effects of global events on supply and demand, and therefore prices. (Heakal, R., 2020)
- "Manufacturing Output" is an economic indicator that measures real output for manufacturing facilities in the United States. This indicator is on a monthly basis in order to bring attention to relevant changes in the production, highlighting structural developments in the economy.
- "Mortgage Market Index" is an indicator on a weekly basis measuring all mortgage applications during the week. This includes all conventional and government applications, all fixed-rate mortgages (FRMs), all adjustable-rate mortgages (ARMs), whether for a purchase or to refinance. (U.S. Mortgage Market Index, 2020)
- "Personal Income" is a monthly report produced by the Bureau of Economic Analysis (BEA) which tracks consumer income that people from wages and salaries, Social Securities and other government benefits, dividends and interest, business ownership, and other sources. It is a good measure of an individual's financial health in the U.S and future consumer spending. Indeed, consumer spending is an important indicator representing a large share of the U.S gross GDP. (Chappelow, J., 2020)

4. Methodology

4.1. Research Question and approach

Based on the discussed concepts and literature review, the research question of this project is: Does the introduction of Alternative Data to the Axyon predictive model produce a more accurate forecast than the original forecasting data?

In order to answer the research question, the alternative data will be integrated into Axyon's machine-learning platform to evaluate results and perform backtesting.

As said before, our project uses Axyon machine-learning platform, the IRIS. This chapter describes, on a high level, the Axyon IRIS model.

4.2. The Predictive Model: Cross-sectional Stock Price Prediction using Deep Learning

The model predictive approach selected uses a so-called pair trading system. It means that given an asset pair (A, B), we want to predict whether A will outperform B or vice versa (cross-sectional prediction), then combine predictions to form a predicted ranking. In this framework, Neural networks return cross-sectional predictions, based on which a ranking is computed. At this point, it is performed a long-short trading strategy meaning going long on the top n assets and short on the bottom n, proportionally to the ranking we previously obtained.

According to the Axyon IRIS process, for both datasets, alternative and standard, comparing feature vectors of two assets (A, B) and trying to predict the relative performance.

This is our target label for the machine learning models. In order to reach this goal, the first idea is to select features since we have a big number of features, and not all of them are useful. This happens thanks to genetic algorithms which is the main technique used in the training part used to discard noisy features. At this point, the models that have the best metrics are selected and trained on the validation set.

4.2.1. Technique: Genetic algorithms

One of the primary processes behind many machine learning applications is feature selection which achieves the goal of finding the most relevant inputs for a model.

This process can be used to discard irrelevant and redundant features that do not add any value to the predictive model.

In this project, a genetic algorithm was used in model training, which is considered a high-level algorithm for feature selection. A Genetic algorithm is a stochastic method for function optimisation inspired by Charles Darwin's theory of natural evolution. "In nature, organisms' genes tend to evolve over successive generations to better adapt to the environment.

Genetic algorithms operate on a population of individuals to produce better and better approximations. At each generation, a new population is created by selecting individuals according to their level of fitness in the problem domain and recombining them together using operators borrowed from natural genetics. The offspring might also undergo mutation." (Gomes, F., Quesada, A. and Lopes, R., 2020)

4.2.2. How does genetic algorithm work?

Five phases are considered in a genetic algorithm:

1. Initialisation
2. Fitness assignment
3. Selection
4. Crossover
5. Mutation

In machine learning and deep learning applications, each individual in the population corresponds to a neural network. In the genetic algorithm, an individual (neural network) presents a set of parameters (variables) known as Genes which are binary values. The number of individuals/neural networks, or population size, depends on the application, while the set of genes represents the whole input variables in the data set.

4.2.3. Initialisation Operator

In this phase, Individuals have to be created and initialised, and usually, this process is made randomly considering that the genetic algorithm is an optimisation method. Genes are joined into a string to form a chromosome (solution).

Having a predicted model represented by a neural network with six possible features, it is possible to generate a population of four individuals which means four neural networks with random features.

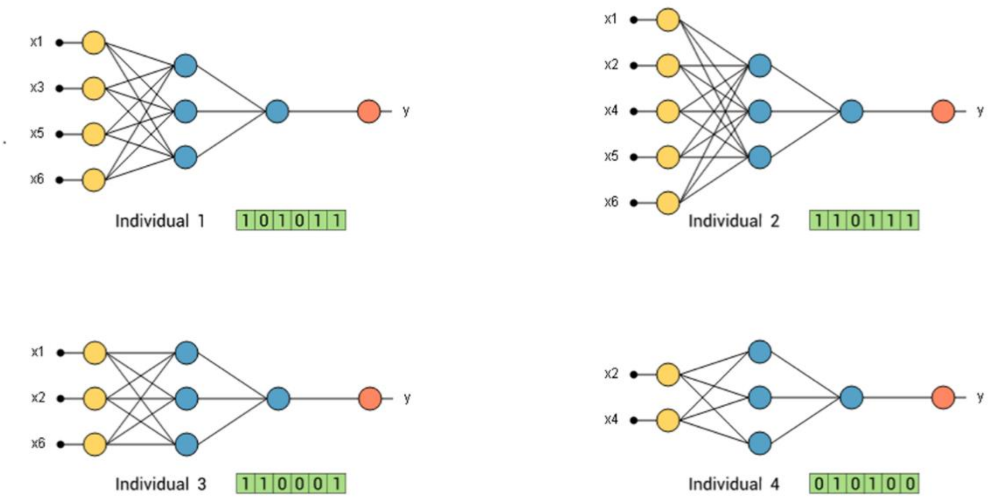


Figure 6. Neural Networks Population (Gomes, F., Quesada, A. and Lopes, R., 2020)

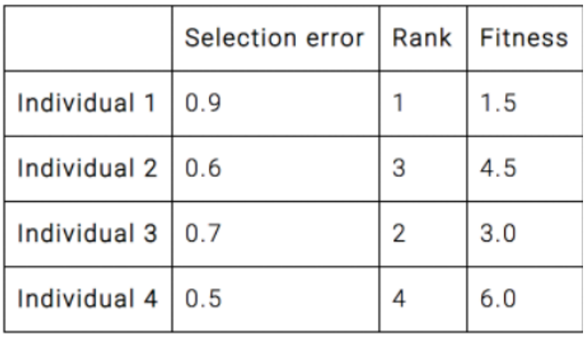
4.2.4. Fitness assignment operator

The fitness function reveals the ability of an individual to compete with other individuals. This happens to assign a fitness score to each individual who represents the probability of an individual to be selected for reproduction. The idea with this operator is to train each neural network with the training instances computing their error with the selection instances.

The bigger is the error of the neural network; the lower will be the fitness score.

There is a well-known method to assign the fitness of each individual called "rank-based fitness assignment" where the selection errors of all the individuals are sorted, and the fitness score of each individual is linked to its position in the ranking system.

On Figure 7, it is possible to find a table related to the example above depicts the selection error, the rank, and the corresponding fitness of each individual.



	Selection error	Rank	Fitness
Individual 1	0.9	1	1.5
Individual 2	0.6	3	4.5
Individual 3	0.7	2	3.0
Individual 4	0.5	4	6.0

Figure 7. Individuals Ranking Table (Gomes, F., Quesada, A. and Lopes, R., 2020)

In the pie chart shown in Figure 8, it is possible to see the area for each individual which is proportional to its fitness score.

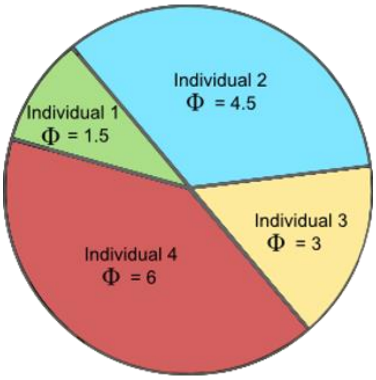


Figure 8. Individuals Fitness Pie (Gomes, F., Quesada, A. and Lopes, R., 2020)

As expected by the table above, the individual 4 presents the most significant area, while individual 1 has the smallest one.

4.2.5. Selection operator

As the name suggests, in this phase, the fittest individuals are selected and their genes, which have shown the capabilities to survive into the environment, are allowed to pass to the next generation. The selection process works firstly according to the level of fitness but then also looking at the population size. Indeed, only half of the population is selected.

There are two main selection methods in this framework, "Elitism selection" which selects for reproduction the individual with the highest fitness, and the "Roulette wheel" where all the individuals are placed on roulette, with areas proportional to their fitness, and once the roulette is turned, the individuals are selected at random. The corresponding individual is chosen for reproduction.

Again, a pie chart in Figure 9 illustrates the selection process with both the methods applied.

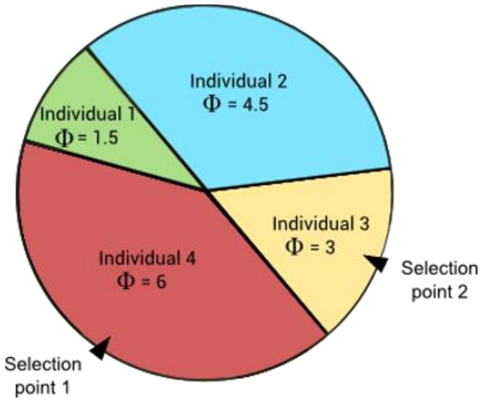


Figure 9. Individuals Selection Process (Gomes, F., Quesada, A. and Lopes, R., 2020)

Applying the elitism method, the neural network with the highest fitness was selected (number 4). In contrast, with the roulette wheel method the individual number 3 is picked even if it presents a lower fitness concerning the individual number 2. This happens according to the stochastic nature of the genetic algorithm.

4.2.6. Crossover operator

Crossover operator is the most significant phase of the genetic algorithm. Having a pair of individuals picked at random, the crossover operator recombines the features within each individual in order to create four offspring representing the new population. From two parents, there are now four offspring. Each of them presents features of both the ancestor and the population size is kept constant concerning the old one.

On Figure 10, it is possible to observe a image which reflects perfectly what was discussed above.

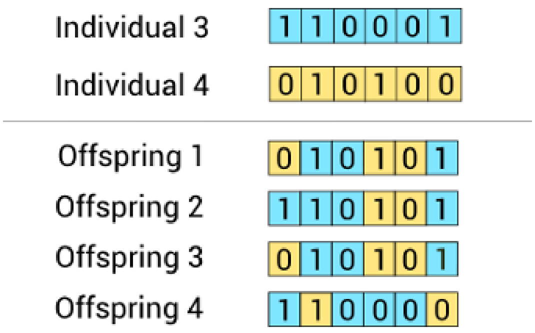


Figure 10. Crossover Method (Gomes, F., Quesada, A. and Lopes, R., 2020)

4.2.7. Mutation operator

The idea with this last phase is to maintain diversity within the population, because it can happen that, in the new population, we have offspring that are very similar to the parents. The mutation operator works to maintain diversity changing the value of some features in the offspring at random. It's important to decide if a feature is mutated or not, and to do this it has to generate a number between 0 and 1. Here, if the number is lower than a value called the mutation rate, the variable is flipped. But, how to choose this mutation rate? Usually, it is chosen to be a portion of 1 (1/m) depending on how many features (m) the network has. Having that value, it's possible to mutate one feature of each network.

Figure11 shows a image that is a representation of the mutation phase of one of the offspring of the new generation, where the fourth input of the neural network mutes and a new population is definitely born.

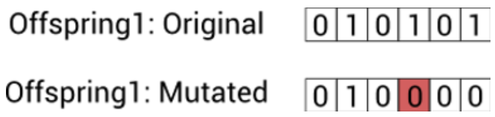


Figure 11. Offspring Mutation Phase (Gomes, F., Quesada, A. and Lopes, R., 2020)

4.2.8. Process and Results of genetic algorithm

The algorithm continues to work and to be repeated until when the population has converged, and it is likely to be more adapted to the environment than the old one and it's also not significantly different from that one. After this, we can say that the genetic algorithm has provided a set of solutions to our problem. The final solution to this process is to find the best individual/neural network in the population, the one that has the lowest value of the selection error.

4.2.9. Pros and cons of genetic algorithm

Considering all the possible techniques that are available in machine learning for feature selection, it's important to select the one which works better in neural network problems. Genetic algorithm is absolutely the most advanced to do that. Indeed, it presents more advantages than disadvantages. Advantages can be summarised in:

- They usually overperform traditional feature selection techniques.
- Genetic algorithms are able to manage data sets with a large number of features
- No previous knowledge of the problem analysed.
- These algorithms can be easily parallelised in computer clusters.

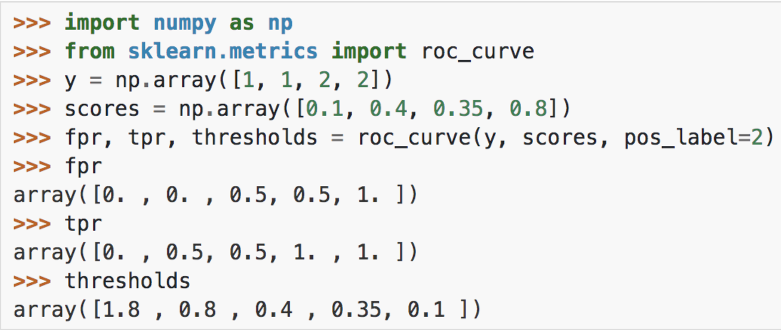
On the other hand, the main disadvantages are the following:

- Genetic Algorithms show relevant computational costs due to the need of training of a model.
- They take a long time to converge due to their stochastic nature. (Gomes, F., Quesada, A. and Lopes, R., 2020)

4.3. Model Evaluation

Applying the genetic algorithm techniques for feature selection explained above into the project, it was obtained essential results from the evaluation of our models. To evaluate our models, we applied the receiver operating characteristic (ROC) curves for individual features to assess their performances.

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of classification model (binary classifier system) at all classification thresholds. It is built by plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate), at various threshold settings. (Machine Learning Crash Course, 2020) Here is a small example of how to use the ROC curve function and a picture in Figure 12 showing an example of such an ROC curve:



```
>>> import numpy as np
>>> from sklearn.metrics import roc_curve
>>> y = np.array([1, 1, 2, 2])
>>> scores = np.array([0.1, 0.4, 0.35, 0.8])
>>> fpr, tpr, thresholds = roc_curve(y, scores, pos_label=2)
>>> fpr
array([0. , 0. , 0.5, 0.5, 1. ])
>>> tpr
array([0. , 0.5, 0.5, 1. , 1. ])
>>> thresholds
array([1.8 , 0.8 , 0.4 , 0.35, 0.1 ])
```

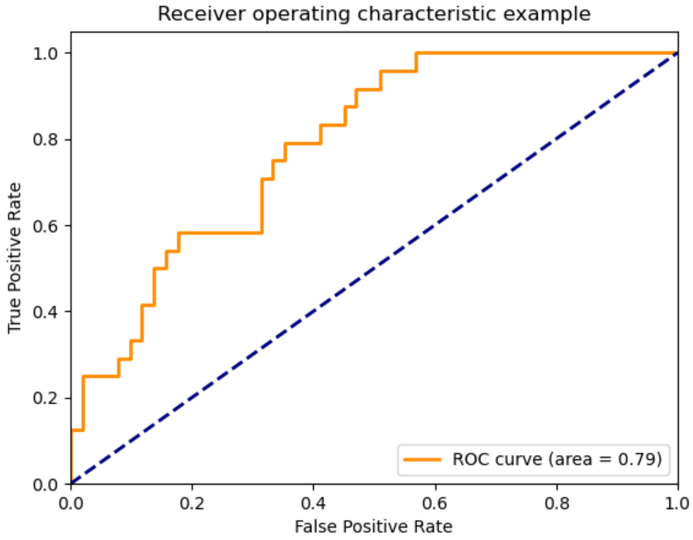


Figure 12. ROC Curve (Pedregosa et al, 2011)

In order to check or visualise the performance of the multiclass classification problem, we use AUC (Area Under the Curve) ROC (Receiver Operating Characteristics) curve which is one of the most crucial evaluation metrics for checking any classification model's performance. "AUC

- ROC curve is a performance measurement for the classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes." (Narkhede, S., 2019)

In the example studied, the area under the receiver operating curve (ROC) is used as a fitness function for the genetic algorithm.

On Figure 13 it is possible to watch the results applying this evaluation method to our standard dataset (without alternative data).

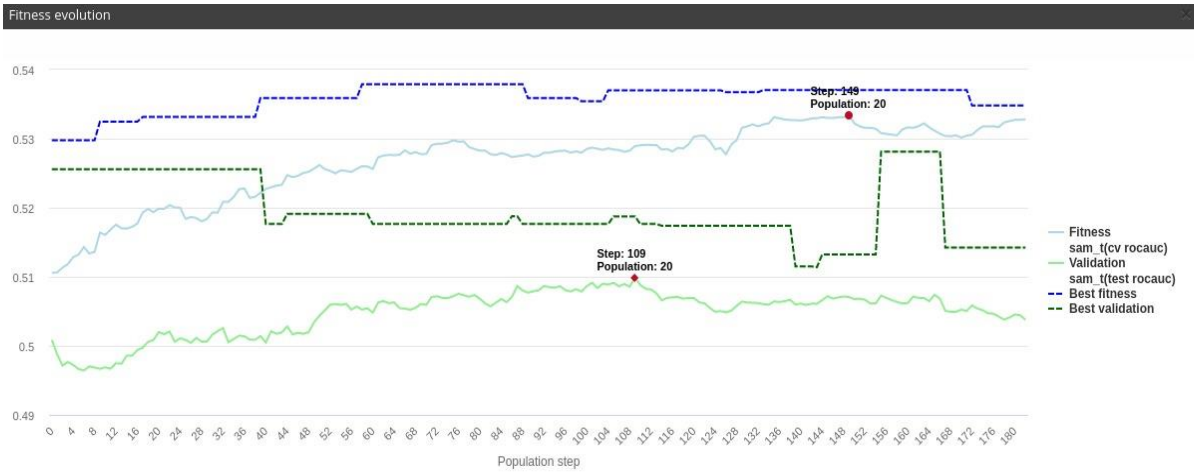


Figure 13. Fitness Evolution Metric – AUC ROC curve (standard dataset)

Thus, on Figure 14 the differences in features performance obtained applying the same method on our alternative dataset.

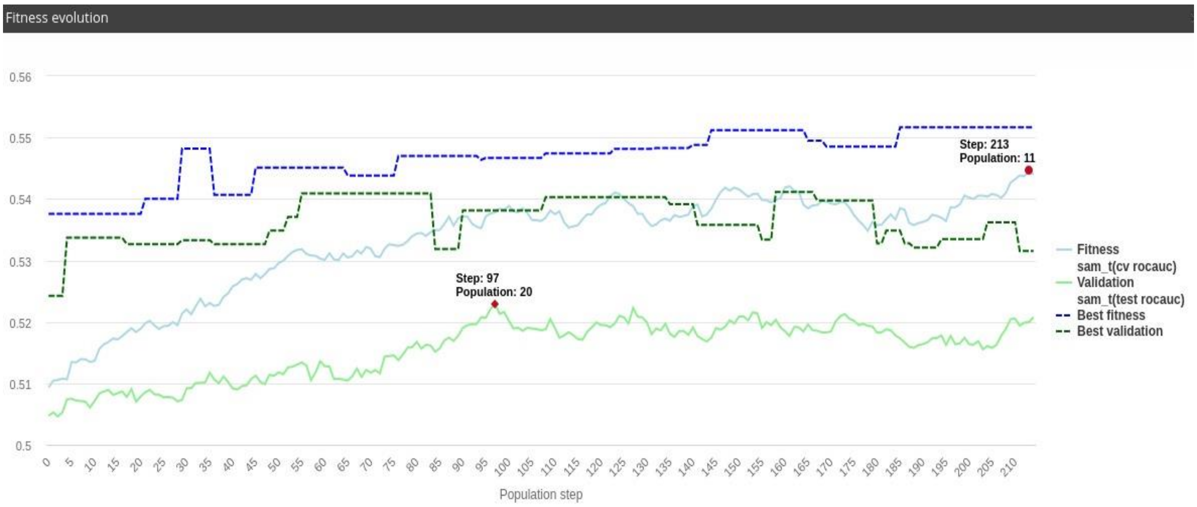
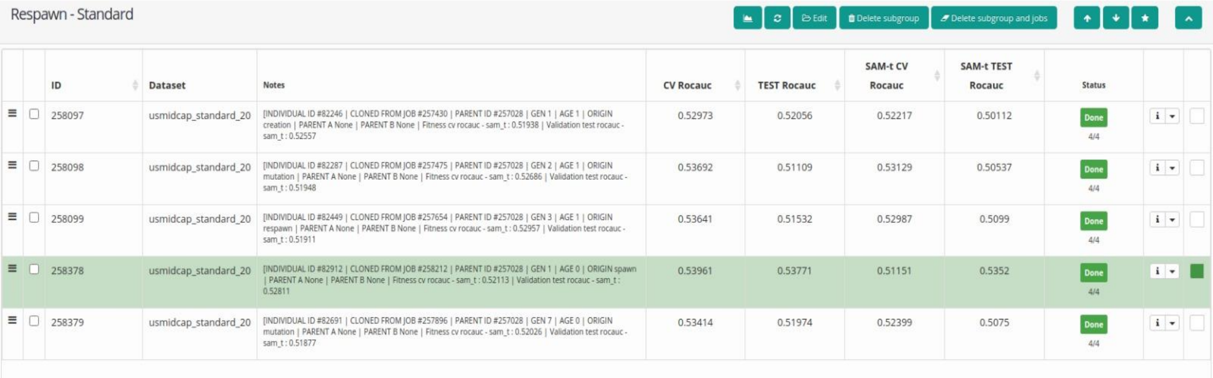


Figure 14. Fitness Evolution Metric – AUC ROC curve (alternative data dataset)

By analysing these two figures above, it can be clearly seen that, applying the ROC AUC scores as model evaluation method on the alternative dataset, performances of the models, which are given by the fitness evolution of those, are significantly better with respect to the ones with the standard dataset.

The goodness of the Alternative dataset concerning the standard dataset is also proven by the differences in the scores assigned applying the ROC AUC method to the best models, which have been selected by the genetic algorithms and belonging to the standard and alternative population.

On Figure 15 a picture showing the best models belonging to the population of our standard dataset, following the ROC AUC scores evaluation method described above. Marked in green, it can be seen the model with the highest score.

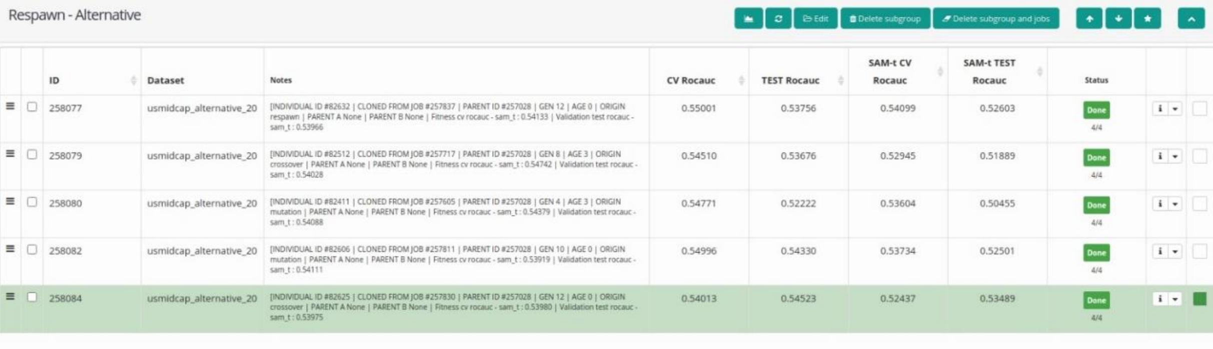


Respawn - Standard

		ID	Dataset	Notes	CV Rocauc	TEST Rocauc	SAM-t CV Rocauc	SAM-t TEST Rocauc	Status		
≡	□	258097	usmidcap_standard_20	[INDIVIDUAL ID #82246 | CLONED FROM JOB #257430 | PARENT ID #257028 | GEN 1 | AGE 1 | ORIGIN creation | PARENT A None | PARENT B None | Fitness cv rocauc - sam_t : 0.51938 | Validation test rocauc - sam_t : 0.52557	0.52973	0.52056	0.52217	0.50112	Done 4/4	i ▾	□
≡	□	258098	usmidcap_standard_20	[INDIVIDUAL ID #82287 | CLONED FROM JOB #257475 | PARENT ID #257028 | GEN 2 | AGE 1 | ORIGIN mutation | PARENT A None | PARENT B None | Fitness cv rocauc - sam_t : 0.52686 | Validation test rocauc - sam_t : 0.51948	0.53692	0.51109	0.53129	0.50537	Done 4/4	i ▾	□
≡	□	258099	usmidcap_standard_20	[INDIVIDUAL ID #82449 | CLONED FROM JOB #257654 | PARENT ID #257028 | GEN 3 | AGE 1 | ORIGIN respawn | PARENT A None | PARENT B None | Fitness cv rocauc - sam_t : 0.52957 | Validation test rocauc - sam_t : 0.51911	0.53641	0.51532	0.52987	0.5099	Done 4/4	i ▾	□
≡	□	258378	usmidcap_standard_20	[INDIVIDUAL ID #82912 | CLONED FROM JOB #258212 | PARENT ID #257028 | GEN 1 | AGE 0 | ORIGIN spawn | PARENT A None | PARENT B None | Fitness cv rocauc - sam_t : 0.52113 | Validation test rocauc - sam_t : 0.52811	0.53961	0.53771	0.51151	0.5352	Done 4/4	i ▾	🟩
≡	□	258379	usmidcap_standard_20	[INDIVIDUAL ID #82691 | CLONED FROM JOB #257896 | PARENT ID #257028 | GEN 7 | AGE 0 | ORIGIN mutation | PARENT A None | PARENT B None | Fitness cv rocauc - sam_t : 0.52026 | Validation test rocauc - sam_t : 0.51877	0.53414	0.51974	0.52399	0.5075	Done 4/4	i ▾	□

Figure 15. Best Models (standard dataset)

On Figure 16 it is possible to observe the picture showing the ROC AUC scoring system applied to the best models belonging to the population of our alternative dataset. As before, marked in green, the model with the highest score on which the backtesting will be launched.



Respawn - Alternative

		ID	Dataset	Notes	CV Rocauc	TEST Rocauc	SAM-t CV Rocauc	SAM-t TEST Rocauc	Status		
≡	□	258077	usmidcap_alternative_20	[INDIVIDUAL ID #82632 | CLONED FROM JOB #257837 | PARENT ID #257028 | GEN 12 | AGE 0 | ORIGIN respawn | PARENT A None | PARENT B None | Fitness cv rocauc - sam_t : 0.54133 | Validation test rocauc - sam_t : 0.53966	0.55001	0.53756	0.54099	0.52603	Done 4/4	i ▾	□
≡	□	258079	usmidcap_alternative_20	[INDIVIDUAL ID #82512 | CLONED FROM JOB #257717 | PARENT ID #257028 | GEN 8 | AGE 3 | ORIGIN crossover | PARENT A None | PARENT B None | Fitness cv rocauc - sam_t : 0.54742 | Validation test rocauc - sam_t : 0.54028	0.54510	0.53676	0.52945	0.51889	Done 4/4	i ▾	□
≡	□	258080	usmidcap_alternative_20	[INDIVIDUAL ID #82411 | CLONED FROM JOB #257605 | PARENT ID #257028 | GEN 4 | AGE 3 | ORIGIN mutation | PARENT A None | PARENT B None | Fitness cv rocauc - sam_t : 0.54379 | Validation test rocauc - sam_t : 0.54088	0.54771	0.52222	0.53604	0.50455	Done 4/4	i ▾	□
≡	□	258082	usmidcap_alternative_20	[INDIVIDUAL ID #82606 | CLONED FROM JOB #257811 | PARENT ID #257028 | GEN 10 | AGE 0 | ORIGIN mutation | PARENT A None | PARENT B None | Fitness cv rocauc - sam_t : 0.53919 | Validation test rocauc - sam_t : 0.54111	0.54996	0.54330	0.53734	0.52501	Done 4/4	i ▾	□
≡	□	258084	usmidcap_alternative_20	[INDIVIDUAL ID #82625 | CLONED FROM JOB #257830 | PARENT ID #257028 | GEN 12 | AGE 0 | ORIGIN crossover | PARENT A None | PARENT B None | Fitness cv rocauc - sam_t : 0.53980 | Validation test rocauc - sam_t : 0.53975	0.54013	0.54523	0.52437	0.53489	Done 4/4	i ▾	🟩

Figure 16. Best Models (alternative data dataset)

By looking at these metrics, it can be seen, as already mentioned before, that the alternative dataset presents a higher score for its models, describing better fitness and performance.

4.4. Model Explainability – SHAP Values

Once the model training is over, it is essential to understand which features have contributed the most to determine predictions and to explain why a particular model has been chosen.

In this framework, we introduce into our project a united approach called SHAP Values which is usually used to explain the output of any machine learning model. Three benefits worth mentioning here. (Dataman, D., 2020).

1. Global interpretability, which is due to the ability of the collective SHAP values to show how much each feature contributes, in a positive or negative way, to the target variable (prediction). Indeed, this benefit allows representing, thanks to the SHAP value plot, the positive or negative impact given by each variable to the target.
2. The second benefit is local interpretability which enables each observation to obtain its own set of SHAP values. This benefit results in an increase in terms of transparency of the factors. This allows understanding the reason behind a particular prediction of a problem and the impacts of the factors selected to achieve the output. So, as the name suggests, not only a global explanation of the model is performed but also a local one, in which the goal is to explain the results made by each individual.
3. The SHAP values can be used for any tree-based model.

Considering as input the features of our best model, which comes from the model training phase, with the SHAP value plot, it is possible to understand the positive and negative impact of the predictors on model output, a pair trading asset performance forecasting.

The code shap.summary_plot(shap_values, X_train) produces the following plot:

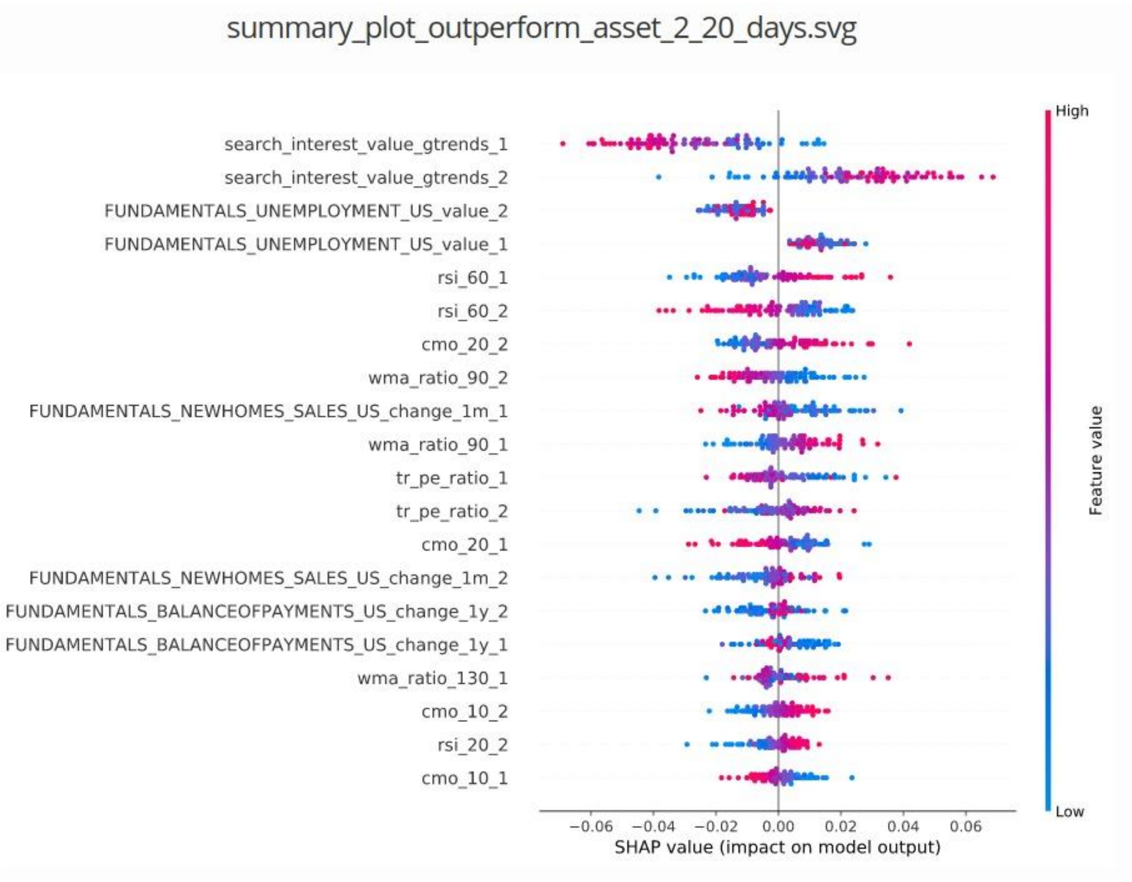


Figure 17. Shap Values - summary plot

On Figure 17 it is shown a plot that is made of all the pair features belonging to the best model obtained at the end of the training phase. It shows the following results:

- Feature importance: Features are showed in descending order according to their importance (contribution).
- Impact: The horizontal location shows whether the effect of that value is associated with a higher or lower prediction.
- Original value: Color shows whether that variable is high (in red) or low (in blue) for that observation.
- Correlation: A high level of the "search_interest_value_gtrends" value (google search trends value) shows a higher and positive contribution of this alternative feature on the model output. The "high" comes from the red colour, and the "positive" impact is shown on the X-axis. Similarly, we will say the "New Homes_Sales" is negatively correlated with the target variable.

5. Results and Discussion

5.1. Backtesting the Model

Backtesting is a crucial component of significant trading system development. It is accomplished by reconstructing, with historical data, trades that would have occurred in the past using rules defined by a given strategy. The result offers statistics to gauge the effectiveness of the strategy." (Kuepper, J., 2020).

Now, with the best models selected, one with alternative data and the other with just standard data, that have the best metrics on the validation set, the backtesting is performed. This technique allows the comparison of the results in order to understand if alternative data has contributed to better performances. The idea with backtesting is to test the results of the output of both the models, simulating what would have happened if we trained our models every 3 months starting from the first portion of the test set used and seeing how it performed in between.

Figure 18 results from the backtesting and shows the cumulative return of our US mid-cap stocks in our best model with alternative features.

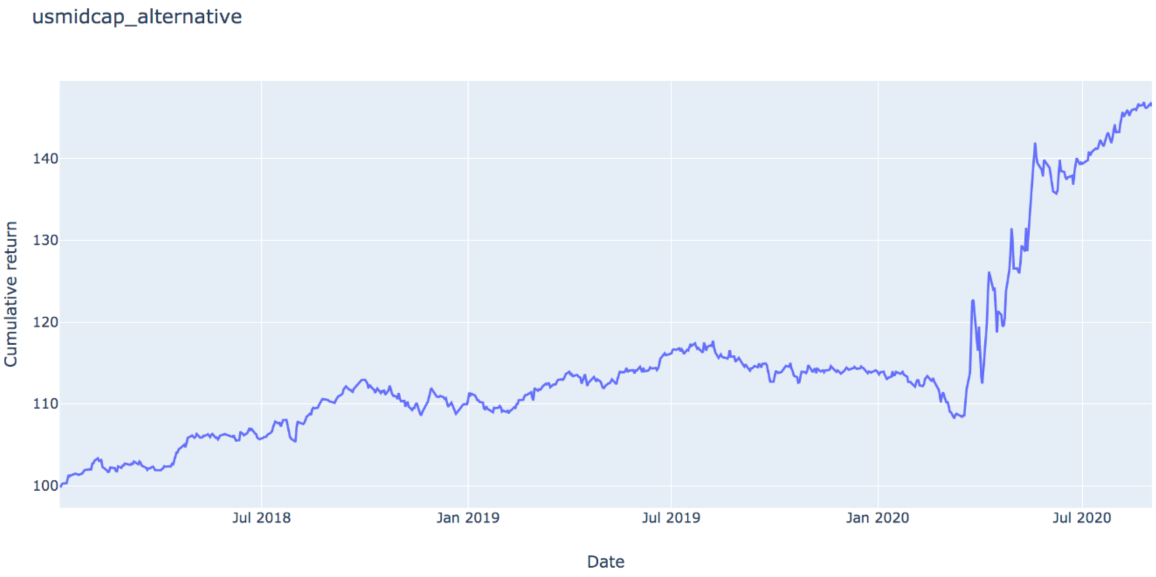


Figure 18. Cumulative Return - Best Model - alternative data dataset

On the other hand, Figure 19 illustrates the cumulative return of the same US mid Cap stocks, although in our best model using our original/standard dataset (without alternative data).

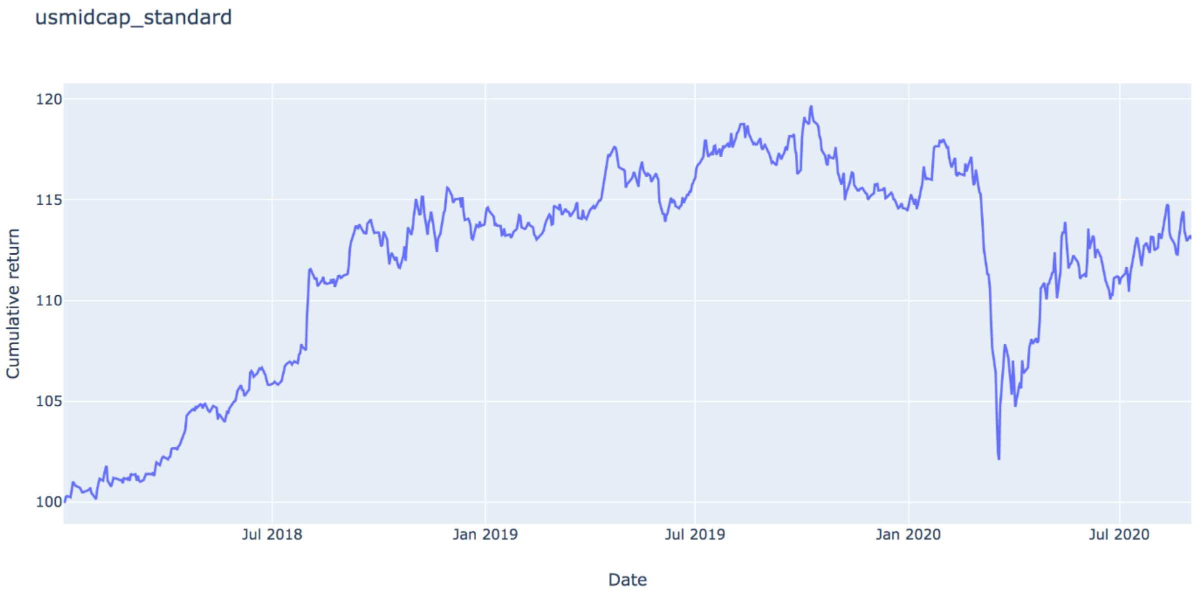
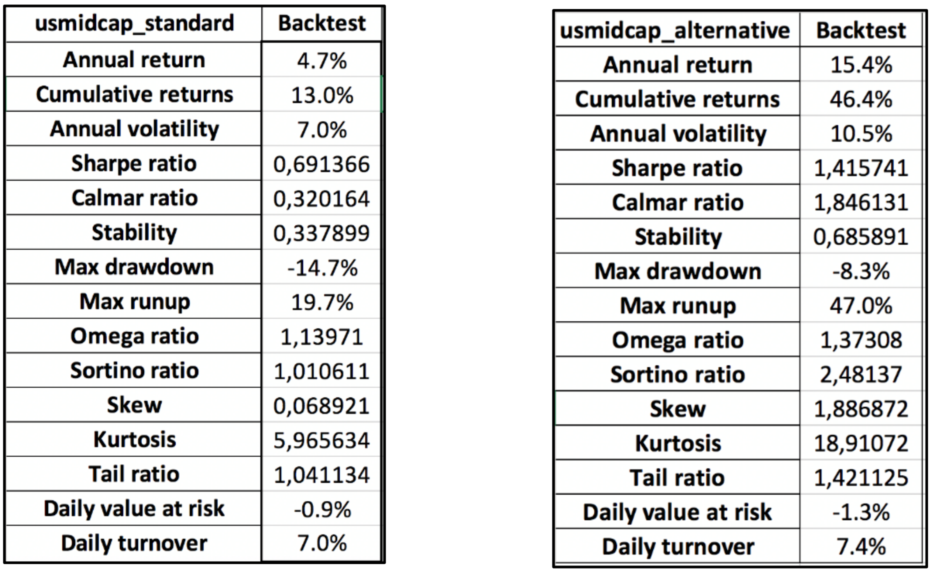


Figure 19. Cumulative Return - Best model - standard dataset

On Figure 20 it is possible to observe two tables showing the differences of the backtest statistic performance results performed on the asset pool between the standard dataset and the one made with the alternative data.



usmidcap_standard	Backtest
Annual return	4.7%
Cumulative returns	13.0%
Annual volatility	7.0%
Sharpe ratio	0,691366
Calmar ratio	0,320164
Stability	0,337899
Max drawdown	-14.7%
Max runup	19.7%
Omega ratio	1,13971
Sortino ratio	1,010611
Skew	0,068921
Kurtosis	5,965634
Tail ratio	1,041134
Daily value at risk	-0.9%
Daily turnover	7.0%

usmidcap_alternative	Backtest
Annual return	15.4%
Cumulative returns	46.4%
Annual volatility	10.5%
Sharpe ratio	1,415741
Calmar ratio	1,846131
Stability	0,685891
Max drawdown	-8.3%
Max runup	47.0%
Omega ratio	1,37308
Sortino ratio	2,48137
Skew	1,886872
Kurtosis	18,91072
Tail ratio	1,421125
Daily value at risk	-1.3%
Daily turnover	7.4%

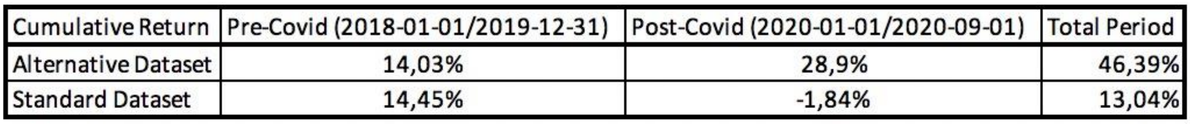
Figure 20. Backtest Statistic Perf. Standard dataset vs Alternative Data dataset

The starting point of our performance analysis is on January 1st 2018, and goes until September 1st 2020, which is the end of the time-frame. It is possible to intuitively conclude that the cumulative return performed on the backtesting trading strategy with the alternative dataset (+46,4%) is much higher than the cumulative return computed with the standard dataset (+13,04%).

Nevertheless, just by looking into the different performances observed on the previous figures between the two datasets, it's possible to highlight a shift in behavior starting in 2020. In order to dive deep into this idea, the analysis was divided into two different dataframes. A pre-COVID 19 time frame and a post-COVID 19 one. A pre-COVID 19 time frame chosen from January 1st 2018 to December 31th 2019, seems to return cumulative returns fairly similar for both datasets. Later, the post-COVID period, from January 1st 2020 to September 1st 2020, reveals a huge gap in terms of cumulative returns.

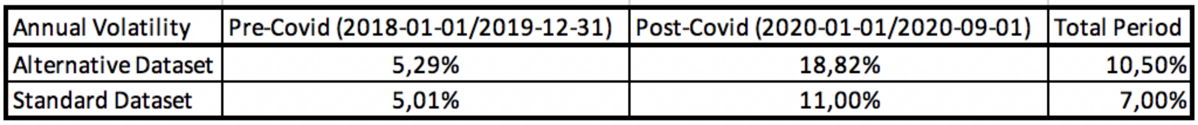
To further analyse this hypothesis, a simple simulation was performed. A new starting point was set, meaning a new initial capital of 100$ at the beginning of 2020, and through the daily returns is has been computed the cumulative return for that period.

Figure 21 and Figure 22 represent two tables summarising the results obtained in terms of cumulative return and annual volatility, splitting the whole period in pre-COVID and post-COVID.



Cumulative Return	Pre-Covid (2018-01-01/2019-12-31)	Post-Covid (2020-01-01/2020-09-01)	Total Period
Alternative Dataset	14,03%	28,9%	46,39%
Standard Dataset	14,45%	-1,84%	13,04%

Figure 21. Cumulative Return; Pre-COVID and Post-COVID analysis



Annual Volatility	Pre-Covid (2018-01-01/2019-12-31)	Post-Covid (2020-01-01/2020-09-01)	Total Period
Alternative Dataset	5,29%	18,82%	10,50%
Standard Dataset	5,01%	11,00%	7,00%

Figure 22. Annual Volatility; Pre-COVID and Post-COVID analysis

On Figure 23 may be observed an image that represents the "underwater plot" of the performances made by the backtesting applied to the standard dataset, which highlights a peak in negative returns from around March 2020.

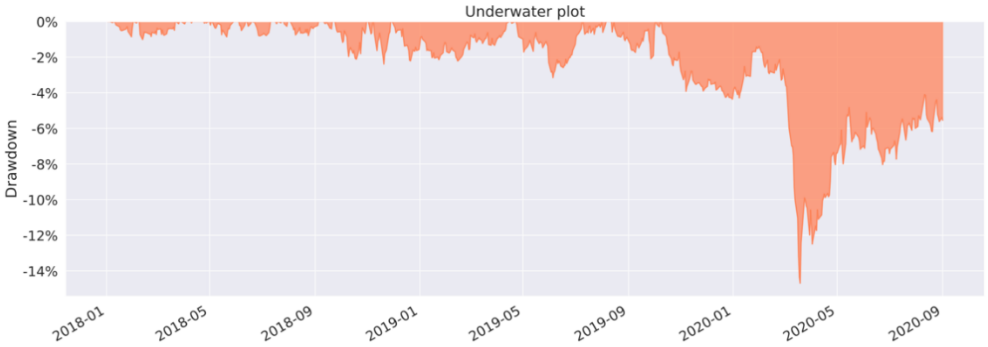


Figure 23. Underwater Plot (standard dataset)

On the other hand, Figure 24. shows the "underwater plot" of the performances made by the backtesting applied to the alternative dataset, also showing the worst performance experienced around the same period as the previous dataset even though with a less significant impact and a convalescence period recovering at a faster pace.

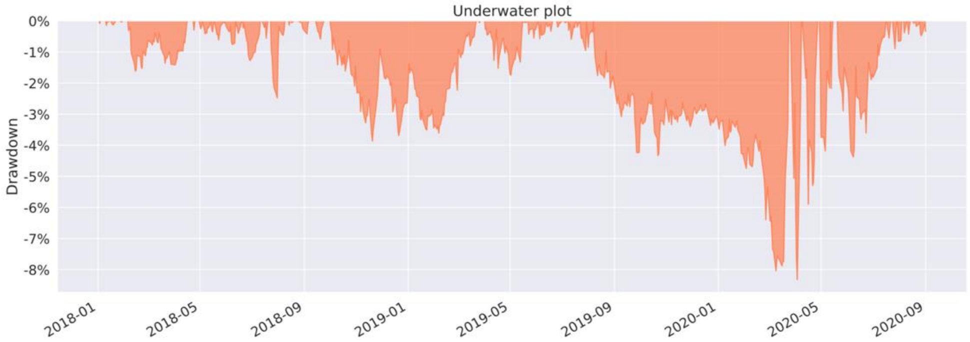


Figure 24. Underwater Plot (alternative dataset)

By looking at those results, it is possible to confirm our thesis of two substantially distinct performances when diving the time-frame into a pre as the post-pandemic scenario. It is also possible to retrain from this analysis that alternative data has contributed to obtain better performances in a critical situation where the whole financial market has experienced a massive sell-off in March 2020. In particular, thanks to alternative data, it has been possible to limit the downside and manage to capture upside in the market. On the contrary, standard data has led to a colossal drawdown in March 2020, putting performances in a position where recovering those losses will be next to impossible.

6. Conclusion

In this project, a study was developed around the possibility of including better and more Alternative Data sources into Axyon ML-platform. Some obstacles were found since the beginning. As alternative data increases its popularity, also the data providers intelligently take advantage of this by setting high prices for the so-called "premium data". In the end, it was possible to obtain Alternative Data without costs associated for the company, due to free data provided by Google and Refinitiv Eikon from Thomson Reuters. The later, it is not a free provider; however, both, the university and the company, currently work with this provider and own a license.

Following the data selection, Axyon AI know-how it was crucial to select the best features and perform the backtesting. Thus, with that computational analysis performed, the results were examined and exhibit some exciting endings. It is possible to generally conclude two different aspects. The first achievement is that, overall, **the alternative data had a positive contribution to the ML-model forecasting performance**. The second one, relies on the fact Alternative Data, **in particular Google Trends data**, had a significantly greater impact on the forecasting performance when compared with its Standard Data-only peer, **specially after the surging of the SARS-CoV-2 novel**, with its most reflected impact dated to March 2020. This latter conclusion was possible due to a closer look into the results. One suggestion to justify this performance is the considerable amount of new people who started investing and trading since the beginning of the COVID-19 pandemic. Perhaps, the inclusion of Alternative Data, such as Google Trends data, allowed the ML-model to better assess where the market was heading, by detecting the behaviour of new investors likely reflected on google searches.

In summary, the underlying project makes an encouraging contribution to reinforcing the importance of the use of Alternative Data sources for ML-based asset performance forecasting, answering positively to our research question. As Emmett Kilduff, CEO and founder of Eagle Alpha told in an article called "What's Alternative Data?" (WSJ, 2020), "You cannot just rely solely on the data sets, but it is definitely a significant advantage."

6.1 Limitations and suggestions for improvement or further research

As previously mentioned in this report, the selection of the data used on this project had some limitations, mainly due to high costs associated with the acquisition of “premium data” or with the absence of historical data, which was an imperative requirement to perform backtesting. Concerning the length of data, if historical data turns out to no longer be criteria for the feature selection, the inclusion of COVID-19 pandemic indicators might improve results significantly.

Another possible limitation found is related to free and easy-access data. Data with these characteristics is available to everybody so its use might start to be mainstreamed on the financial industry. Matt Levine on his *Bloomberg* newsletter “Money Stuff” found three lessons on this topic, first, if someone can buy better or faster data, they will have a significant advantage and can expect juicy profits. Second, if the better or faster data can be purchased or used (in case it's free) by anyone, then everybody must do it:“If you don’t, then you have a big disadvantage compared to the people who do, and you can lose a lot of money.” Third, either way, whoever can sell better or faster data will make a lot of money.

Appendix

```python
import sys
from datetime import datetime, timedelta, date
import time

from tqdm import tqdm
import numpy as np
import pandas as pd
from pytrends.request import TrendReq
from pytrends.exceptions import ResponseError

from lib.db.dbmanager_mysql import FastDBManager

def _fetch_data(trendreq, kw_list, timeframe='today 3-m', cat=0, geo='', gprop='') -> pd.DataFrame:
    # TODO: check consistency and improve, set correct timezone etc
    """Download google trends data using pytrends TrendReq and retries in case of a ResponseError."""
    attempts, fetched = 0, False
    while not fetched:
        try:
            trendreq.build_payload(kw_list=kw_list, timeframe=timeframe, cat=cat, geo=geo, gprop=gprop)
        except ResponseError as err:
            print(err)
            print(f'Failed {kw_list}, Trying again in {60 + 5 * attempts} seconds.')
            time.sleep(60 + 5 * attempts)
            attempts += 1
            if attempts > 3:
                print('Failed after 3 attemps, abort fetching.')
                break
        else:
            fetched = True
    iot = trendreq.interest_over_time()
    if not iot.empty:
        iot.drop(columns=['isPartial'], inplace=True)
    return iot


def get_daily_trend(trendreq, keywords, current_date, delta=269, sleep=5, tz=0, verbose=False):
    """Stich and scale consecutive daily trends data between start and end date.
    This function will first download piece-wise google trends data and then
    scale each piece using the overlapped period.
        Parameters
        ----------
        trendreq : TrendReq
            a pytrends TrendReq object
        keyword: str
            currently only support single keyword, without bracket
        start: str
            starting date in string format:YYYY-MM-DD (e.g.2017-02-19)
        end: str
            ending date in string format:YYYY-MM-DD (e.g.2017-02-19)
        cat, geo, gprop, sleep:
            same as defined in pytrends
        delta: int
            The length(days) of each timeframe fragment for fetching google trends data,
            need to be <269 in order to obtain daily data.
        overlap: int
            The length(days) of the overlap period used for scaling/normalization
        tz: int
            The timezone shift in minute relative to the UTC+0 (google trends default).
            For example, correcting for UTC+8 is 480, and UTC-6 is -360
    """
    init_end_d = end_d = datetime.strptime(current_date, '%Y-%m-%d') if type(current_date) == str else current_date
    init_end_d.replace(hour=23, minute=59, second=59)
    assert delta < 269
    delta = timedelta(days=delta)

    start_d = end_d - delta

    df = pd.DataFrame()

    tf = start_d.strftime('%Y-%m-%d')+' '+end_d.strftime('%Y-%m-%d')
    if verbose: print('Fetching \''+keywords+'\' for period:'+tf)
    df = _fetch_data(trendreq, keywords, timeframe=tf)
    if df.empty:
        return df

    # in case of short query interval getting banned by server
    time.sleep(sleep)

    df.sort_index(inplace=True)
    #The daily trend data is missing the most recent 3-days data, need to complete with hourly data
    if df.index.max() < init_end_d :
        tf = 'now 7-d'
        hourly = _fetch_data(trendreq, keywords, timeframe=tf)

        #convert hourly data to daily data
        daily = hourly.groupby(hourly.index.date).sum()

        #check whether the first day data is complete (i.e. has 24 hours)
        daily['hours'] = hourly.groupby(hourly.index.date).count()
        if daily.iloc[0].loc['hours'] != 24: daily.drop(daily.index[0], inplace=True)
        daily.drop(columns='hours', inplace=True)

        daily.set_index(pd.DatetimeIndex(daily.index), inplace=True)

        # find the overlapping date
        intersect = df.index.intersection(daily.index)
        if verbose: print('Normalize by overlapping period:'+(intersect.min().strftime('%Y-%m-%d'))+' '+(intersect.max().strftime('%Y-%m-%d')))
        # scaling use the overlapped today-4 to today-7 data
        coef = df.loc[intersect].iloc[:,0].max() / daily.loc[intersect].iloc[:,0].max()
        daily = (daily*coef).round(decimals=0)
        df = pd.concat([daily, df], axis=1)

    # taking averages for overlapped period
    df = df.mean(axis=1)
    # # Correct the timezone difference
    df.index = df.index + timedelta(minutes=tz)
    df = df[start_d:init_end_d]

    return df

class GTrendManager():
    def __init__(self, instrument_codes, logger, ref_key_enabled=False):
        self.dbmanager = FastDBManager()
        self.logger = logger
        self.instrument_keys_df = self._load_search_keys(instrument_codes, ref_key_enabled)

        self.ref_key_enabled = ref_key_enabled

    def _load_search_keys(self, instrument_codes, reference_key_enabled=False):
        if reference_key_enabled:
            raise NotImplementedError("TODO: implement the use of a reference key to put all search interests on the "\
                                      "same scale")
        instrument_list = "','".join(instrument_codes)
        instrument_where = f"('{instrument_list}')"
        query = f"SELECT * FROM sn_search_keys WHERE instrument_code in {instrument_where};"
        search_key_table = self.dbmanager.load_dataframe(query)

        return search_key_table

    def _download_daily_data(self, date_from, date_to, update_data):
        initial_date = datetime.strptime(date_from, "%Y-%m-%d")
        if datetime.date(initial_date) == date.today():
            initial_date = initial_date - timedelta(days=1)
        last_date = datetime.strptime(date_to, "%Y-%m-%d")
        if datetime.date(last_date) == date.today():
            last_date = last_date - timedelta(days=1)
        estimated_points = np.busday_count(datetime.date(initial_date), datetime.date(last_date))

        # TODO: set correctly date, timezone etc
        pytrend = TrendReq(hl='en-US')

        pbar = tqdm(total=estimated_points*len(self.instrument_keys_df), file=sys.stdout)
        # Iterate dates
        current_date = initial_date
        while current_date <= last_date:
            # Iterate instruments
            for _, row in self.instrument_keys_df.iterrows():
                search_terms = [row.search_key]
                if self.ref_key_enabled:
                    search_terms.append(row.reference_key)

                series = get_daily_trend(pytrend, search_terms, current_date, 90)
                if series.empty:
                    continue

                today_value = series.iloc[-1]/100
                daily_change = (series.iloc[-1] - series.iloc[-2])/100
                vs_7d_avg = (series.iloc[-1] - series.iloc[-7:].mean())/100
                vs_30d_avg = (series.iloc[-1] - series.iloc[-30:].mean())/100
                vs_90d_avg = (series.iloc[-1] - series.mean())/100

                query = f"INSERT INTO sn_search_interest_daily (date, instrument_code,  search_interest_value, " \
                        f"daily_change, vs_7d_avg, vs_30d_avg, vs_90d_avg) VALUES ('{current_date}', '{row.instrument_code}', " \
                        f"{today_value:.2f}, {daily_change:.2f}, {vs_7d_avg:.2f}, {vs_30d_avg:.2f}, {vs_90d_avg:.2f})"
                if update_data:
                    query += f" ON DUPLICATE KEY UPDATE search_interest_value={today_value:.2f}, daily_change={daily_change:.2f}, "\
                        f"vs_7d_avg={vs_7d_avg:.2f}, vs_30d_avg={vs_30d_avg:.2f}, vs_90d_avg={vs_90d_avg:.2f}"
                self.dbmanager.execute_query(query)
                pbar.update(1)
            current_date += timedelta(days=1)
        return

    def _download_monthly_data(self, date_from, date_to, update_data):
        # TODO: set correctly date, timezone etc
        pytrend = TrendReq(hl='en-US')

        # Iterate instruments
        for _, row in self.instrument_keys_df.iterrows():
            search_terms = [row.search_key]
            if self.ref_key_enabled:
                search_terms.append(row.reference_key)

            timeframe = f"{date_from} {date_to}"
            series = _fetch_data(pytrend, search_terms, timeframe)
            if series.empty:
                continue

            col_names = ["search_interest_value", "monthly_change", "vs_3m_avg", "vs_6m_avg", "vs_12m_avg"]
            df = pd.DataFrame(columns=col_names, index=series.index)
            df["search_interest_value"] = series/100
            df["monthly_change"] = series.diff(1)/100
            df["monthly_change"] = df["monthly_change"].where(pd.notnull(df["monthly_change"]), 0)
            df["vs_3m_avg"] = series / series.rolling(window=3).mean()
            df["vs_3m_avg"] = df["vs_3m_avg"].where(pd.notnull(df["vs_3m_avg"]), 1)
            df["vs_6m_avg"] = series / series.rolling(window=6).mean()
            df["vs_6m_avg"] = df["vs_6m_avg"].where(pd.notnull(df["vs_6m_avg"]), 1)
            df["vs_12m_avg"] = series / series.rolling(window=12).mean()
            df["vs_12m_avg"] = df["vs_12m_avg"].where(pd.notnull(df["vs_12m_avg"]), 1)

            for date, entry in df.iterrows():
                query = f"INSERT INTO sn_search_interest_monthly (date, instrument_code,  search_interest_value, " \
                        f"monthly_change, vs_3m_avg, vs_6m_avg, vs_12m_avg) VALUES ('{date}', '{row.instrument_code}', " \
                        f"{entry.search_interest_value:.2f}, {entry.monthly_change:.2f}, " \
                        f"{entry.vs_3m_avg:.2f}, {entry.vs_6m_avg:.2f}, {entry.vs_12m_avg:.2f})"
                if update_data:
                    query += f" ON DUPLICATE KEY UPDATE search_interest_value={entry.search_interest_value:.2f}, " \
                        f"monthly_change={entry.monthly_change:.2f}, vs_3m_avg={entry.vs_3m_avg:.2f}, "\
                        f"vs_6m_avg={entry.vs_6m_avg:.2f}, vs_12m_avg={entry.vs_12m_avg:.2f}"
                self.dbmanager.execute_query(query)
        return

    def download_trends(self, date_from, date_to, frequency, update_data=True):
        if frequency == "daily":
            self._download_daily_data(date_from, date_to, update_data)
        elif frequency == "monthly":
            self._download_monthly_data(date_from, date_to, update_data)
        self.logger.info("Download completed")


```

References

Balasubramaniam, K. (2020, August 28th). What impact does a higher non-farm payroll have on the forex market? Retrieved from https://www.investopedia.com/ask/answers/06/nonfarmpayrollandforex.asp

Chappelow, J. (2020, September 16th). Personal Income And Outlays Definition. Retrieved from https://www.investopedia.com/terms/p/personal-income-outlays.asp

Chen, J. (2020, August 28th). Consumer Price Index (CPI) Definition. Retrieved from https://www.investopedia.com/terms/c/consumerpriceindex.asp

Dataman, D. (2020). *Explain Your Model with the SHAP Values*. Retrieved from https://towardsdatascience.com/explain-your-model-with-the-shap-values-bc36aac4de3d

Chrisley, R., & Begeer, S. (2000). *Artificial intelligence: Critical concepts*. London: Routledge.

Constable, S. (2019). What Is Alternative Data? *Wall Street Journal.* Retrieved from https://www.wsj.com/articles/what-is-alternative-data-11575860400

Otani, A., & Wursthorn M. (2020). Wall Street's New Metrics. *Wall Street Journal.* Retrieved from https://www.wsj.com/articles/restaurant-reservations-driving-directions-and-other-indicators-wall-street-is-watching-11592740801

Dataiku. (n.d.). Alternative Data in Financial Markets. Retrieved from https://pages.dataiku.com/white-paper-alternative-data-in-financial-markets

Fernando, J. (2020, September 16th). New Home Sales Definition. Retrieved from https://www.investopedia.com/terms/n/newhomesales.asp

Gomes, F., Quesada, A. and Lopes, R. (2020) *Genetic algorithms for feature selection.* Retrieved from https://www.neuraldesigner.com/blog/genetic_algorithms_for_feature_selection#:~:text=One of the most advanced,better adapt to the environment

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. 1-26; 96-161; 164-223.

Google Developers, *Machine Learning Crash Course* Retrieved from https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Heakal, R. (2020, August 28th). What Is International Trade? Retrieved from https://www.investopedia.com/insights/what-is-international-trade/

Kagan, J. (2020, September 09th). Existing Home Sales Definition. Retrieved from https://www.investopedia.com/terms/e/existinghomesales.asp

Kagan, J. (2020, September 16th). Initial Claims. Retrieved from https://www.investopedia.com/terms/i/initialclaims.asp

Kolanovic, M. & R. Krishnamachari (2017). Big data and AI strategies: Machine learning and alternative data approach to investing. *White paper, JP Morgan, Quantitative and Derivatives Strategy.*

Kuepper, J. (2020, August 28th). The Importance of Backtesting Trading Strategies. Retrieved from https://www.investopedia.com/articles/trading/05/030205.asp

Narkhede, S. (2019). *Understanding AUC - ROC Curve*. Retrieved from https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Levine, M. (2020) Do We Really Have to Talk About TikTok Key Money? *Bloomberg Opinion*. Retrieved from https://www.bloomberg.com/opinion/articles/2020-08-04/trump-on-tiktok-sale-payment-do-we-really-have-to-talk-about-it

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *The Journal of machine Learning research*, *12*, 2825-2830.

Prado, M. L. (2018). *Advances in financial machine learning*. Hoboken, New Jersey: Wiley.

Ramos, E. G. and Martínez, F. V. (2013). A review of artificial neural networks: How well do they perform in forecasting time series? *Journal of Statistical Analysis*, 6(2).

Russell, S.J., Norvig, P. (1995). Artificial Intelligence: A Modern approach. New Jersey: Prentice Hall, Englewood Cliffs.

Sarle, W. S.(1994)" Neural Networks and Statistical Models".

Smith, N. (2020). Alternative Data Offers a Lot. Just be careful. Retrieved from https://www.bloomberg.com/opinion/articles/2020-07-15/coronavirus-economic-turmoil-makes-case-for-alternative-data

U.S. Census Bureau (2020). *United States Core Retail Sales*. Retrieved from https://www.census.gov/retail/index.html

U.S. Mortgage Market Index (2020). Retrieved from https://www.investing.com/economic-calendar/mortgage-market-index-1427

Williams, W. (2020, September 04th). Unemployment Rate by State. Retrieved from https://www.investopedia.com/unemployment-rate-by-state-4843541

World Economic Forum. (2018). The New Physics of Financial Services. Understanding How Artificial Intelligence Is Transforming the Financial Ecosystem. Retrieved from https://www2.deloitte.com/uk/en/pages/financial-services/articles/artificial-intelligence-transforming-financial-ecosystem-deloitte-fsi.html?id=gb:2or:3vu:4WEFAI:5eng:6fs:short