| FULL LEGAL NAME | **LOCATION (COUNTRY)** | **EMAIL ADDRESS** | **MARK X FOR ANY NON-CONTRIBUTING MEMBER** |
| --- | --- | --- | --- |
| Haowen Shen | China | 0xconnor3@gmail.com |  |
| Justine Otieno Kitoto |  |  | x |
| siddharth dixit |  |  | X |

|  |  |
| --- | --- |
| **Statement of integrity:** By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an “X” above). | |
| **Team member 1** | Haowen Shen |
| **Team member 2** |  |
| **Team member 3** |  |

|  |
| --- |
| Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed.  **Note:** You may be required to provide proof of your outreach to non-contributing members upon request. |
| siddharth dixit never reached out to us |

## Q1. Data Types Used for Stock Market Movement Prediction

### Primary Data Sources and Collection

The researchers used historical price data for three Exchange-Traded Funds (ETFs) obtained from Yahoo Finance:

* iShares MSCI Chile ETF (ECH) - an emerging market ETF
* iShares MSCI Brazil ETF (EWZ) - an emerging market ETF
* iShares Core S&P 500 ETF (IVV) - a developed market ETF for comparison

The data covered the period from December 12, 2009, to January 1, 2020. This timeframe was deliberately selected to exclude market distortions caused by the COVID-19 pandemic, focusing on more typical market behavior patterns.

### Raw Data Components

For each ETF, the following daily price data points were collected:

* Opening price (Open)
* Highest price (High)
* Lowest price (Low)
* Closing price (Close)
* Trading volume (Volume)
* Adjusted closing price (Adjusted Close) - adjusted for splits, dividends, and capital gain distributions

### Technical Indicator Derivation

From these six fundamental data points, the researchers expanded their feature set dramatically:

* They used the Pandas Technical Analysis Library (Pandas TA) to calculate 210 additional technical indicators
* Combined with the 6 original features, this created a total of 216 daily features for analysis
* These technical indicators fall into various categories including: Candles, Cycles, Momentum, Overlap, Performance, Statistics, Trend, Utility, Volatility, and Volume

### Target Variable Construction

The researchers created a binary classification target variable (Γ) defined as:

* Γ(t) = 1 if Open(t) - Open(t-1) > 0 (price increased)
* Γ(t) = -1 otherwise (price decreased or remained the same)

This transformed the prediction task into a binary classification problem predicting whether the opening price would rise or not on the following day.

## Importance of Technical Indicators in Forecasting Stock Price Trends

### Addressing Market Complexity

The paper emphasizes that stock market prediction is challenging due to the "energetic, nonlinear, nonparametric, and chaotic properties of stock information." Technical indicators help address this complexity by:

* Capturing patterns that might not be evident in raw price data
* Providing quantitative measurements of market psychology and momentum
* Revealing different aspects of market behavior across various timeframes

## Q2. Security Understanding

The iShares MSCI Chile ETF (ECH) is an exchange-traded fund that tracks the performance of the Chilean equity market by replicating an index composed of Chilean stocks. As an emerging market ETF, it provides investors with exposure to the Chilean economy without requiring direct investment in individual Chilean securities. ECH is managed by BlackRock, one of the world's largest asset management companies.

## Asset Type and Composition

ECH is composed primarily of large and mid-capitalization Chilean equities. According to the research paper, the fund has significant sector concentration with the following distribution:

* Financials: 21.53%
* Materials: 21.28%
* Utilities: 18.73%
* Consumer Staples: 13.92%
* Energy: 8.34%

## Price History and Performance

The price history of ECH from 2009 to 2020 (the period analyzed in the study) shows considerable volatility typical of emerging markets:

* Minimum price: $29.30
* Maximum price: $80.25
* Median price: $46.48
* Mean price: $50.10

## Trading Statistics

The study reports the following statistical measures for ECH's opening price:

* 1st Quartile: $40.35
* 3rd Quartile: $59.84
* Price Range: $50.95 (from min to max)

## Why Classification Rather Than Regression?

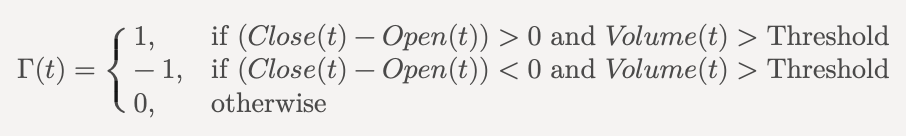
1. **Directional focus**: The paper emphasizes that for many investors, the direction of price movement (up or down) is more important than the exact magnitude of the change. Classification provides this binary signal directly.
2. **Reduced complexity**: Stock prices are highly volatile and affected by numerous factors, making precise numerical predictions extremely challenging. By simplifying to directional prediction, the model becomes more robust.
3. **Performance evaluation**: Classification accuracy is easier to evaluate than regression errors when dealing with highly volatile time series data.
4. **Practical application**: For trading strategies, especially in emerging markets with higher volatility, knowing the direction often provides sufficient information for decision-making.

## Alternative Classification Definitions

The authors defined their classification variable Γ based on whether the opening price increased from the previous day. Two alternative approaches could be:

1. **Volatility-adjusted price movement**: Γ(t) = 1 if (Close(t) - Open(t))/σ > threshold Γ(t) = -1 otherwise

Where σ represents the standard deviation of price changes over a rolling window. This would classify days based on whether the price movement exceeds typical volatility, identifying more significant moves and filtering out noise.

1. **Volume-weighted price direction:**  This classification would give more weight to price movements accompanied by higher trading volumes, potentially identifying more significant market moves with stronger conviction, as higher volume often indicates stronger market sentiment behind a price movement.

**Q3. Methodology Understanding**

The original "Materials and Methods" section can be separated into two distinct sections. Here's how Section 2 (Data) should be organized:

### 2. Data

#### 2.1 Stocks Analyzed

* Selection of ETFs (ECH, EWZ, IVV)
* Time frame rationale (December 12, 2009, to January 1, 2020)
* Market exposure characteristics
* Pre-pandemic data importance for feature selection

#### 2.2 Data Sources

* Yahoo Finance data components (Open, High, Low, Close, Volume, Adjusted Close)
* Descriptive statistics of the opening price data

#### 2.3 Data Processing

* Definition of the target variable Γ (class label for prediction)
* Data normalization using min-max approach
* Data cleaning (handling missing values)
* Class assignment methodology

#### 2.4 Technical Indicators

* Implementation of Pandas Technical Analysis Library
* Extension of original 6 attributes to 216 daily features
* Types of indicators calculated (momentum, trend, volatility, etc.)

## New Section 3: Methodology

The methodology section should focus on the analytical approaches and techniques used:

### 3. Methodology

#### 3.1 Cross-Industry Standard Process for Data Mining (CRISP-DM)

* Six-stage methodological framework
* Implementation in the ETF prediction context

#### 3.2 Feature Selection Methods

* Low Variance approach
* Chi-Squared technique
* LASSO (Least Absolute Shrinkage and Selection Operator)
* Tree-based Feature Selection
* Principal Feature Analysis
* Mean Absolute Difference (MAD)
* Dispersion Ratio (DR)

#### 3.3 Neural Network Model

* Multilayer Perceptron (MLP) architecture
* Configuration parameters for prediction
* Training process details

#### 3.4 Cross-validation

* K-fold cross-validation process (K=10)
* Algorithm implementation for validation

#### 3.5 Experimental Design

* Comprehensive experimental methodology (Algorithm 2)
* Feature selection optimization process
* Model evaluation approach

### Dividing Descriptive Statistics from Models

The distinction between descriptive statistics and models could be made as follows:

**Descriptive Statistics:**

* Pearson's Correlation (measures relationships between variables)
* Mean Absolute Difference (describes data dispersion)
* Dispersion Ratio (measures feature relevance)
* Low Variance approach (identifies constant features)

**Models:**

* LASSO (a regularization technique for feature selection and model fitting)
* Tree-based Feature Selection (machine learning approach for feature importance)
* Principal Feature Analysis (dimensionality reduction technique)
* Chi-Squared (statistical test for independence)
* Neural Network Model (predictive model)

### Technical Indicators Optimization Process

The optimization process for technical indicators in the paper follows these steps:

1. **Feature Calculation and Preprocessing**: The authors calculate 216 technical indicators for each ETF, normalize the data, and clean it by removing days with missing values.
2. **Statistical Feature Selection**: Multiple statistical measures are applied to identify salient features, including correlation analysis, MAD, DR, and other techniques.
3. **Feature Subset Creation**: The algorithm identifies features that appear in at least a specified number of subsets defined by statistical measures, creating different Selected(n) subsets.
4. **Iterative Evaluation**: The MLP model is trained and evaluated on different subsets of features using K-fold cross-validation (K=10).
5. **Performance Measurement**: Average accuracy across cross-validation folds is calculated for each subset of features.

The importance of this optimization process is highlighted in the results, where the Selected(5) subset achieves better prediction accuracy (77.82-80.27%) while using only 4.16-5.09% of the original features. This significantly reduces computational resources and training time (by 84.68% on average) while improving accuracy by approximately 13.63%.

The optimization is particularly valuable because:

* It identifies the most relevant features for each specific ETF
* It removes redundant information that could lead to overfitting
* It creates more efficient models with fewer parameters
* It improves generalization by focusing on the most predictive indicators
* It significantly reduces computational costs while maintaining or improving performance

**Step 1: Financial Problem**

**The Financial Problem the Authors Aim to Solve**

Based on Section 3 of the paper, the authors aim to address the challenge of predicting ETF (Exchange-Traded Fund) trend directions in emerging markets with greater accuracy and computational efficiency. Specifically, they focus on:

1. Identifying the most salient technical indicators that effectively predict ETF price movements in emerging markets
2. Reducing computational costs while maintaining or improving prediction accuracy
3. Providing investors with a reliable tool for market timing and risk management in emerging markets
4. Enabling better decision-making for entering or exiting positions in emerging market ETFs

As stated in Section 3.1 (Practical Implications for Investors in Emerging Markets), their model helps investors with "market timing and risk management within emerging markets" and allows them to "ascertain the optimal timing for entering or exiting positions in ETFs, effectively maximizing returns and minimizing losses."

**Differences Between Predicting Emerging vs. Developed Markets**

The paper reveals several key differences between predicting stock movements in emerging versus developed markets:

1. **Different Technical Indicators**: The analysis shows that emerging markets depend more on cyclical behaviors of prices, while developed markets do not exhibit this same dependence. This is evident in the feature selection results where "Cycles" category features appear in both emerging market ETFs (ECH and EWZ) but not in the developed market ETF (IVV).
2. **Quantitative vs. Qualitative Features**: The authors note that "emerging markets' predictions use quantitative features, while developed markets rely more on qualitative features." For example, TTM Trend (a qualitative indicator) is only selected for the developed market ETF (IVV).
3. **Volume Indicators**: Different volume indicators are relevant for each market type. For emerging markets, Archer's On Balance Volume (AOBV) is selected, while Price Volume Rank (PVR) is selected for developed markets. These indicators measure volume-price relationships differently.
4. **Market Exposure Differences**: The paper suggests that market exposure (sector distribution) influences which indicators are most predictive. Table 1 shows significantly different sector distributions between emerging market ETFs and the developed market ETF.
5. **Volatility Patterns**: Emerging markets exhibit higher volatility and different risk characteristics, requiring specialized feature selection for effective prediction.

This distinction is significant for the model's design because it demonstrates that a one-size-fits-all approach to technical analysis is suboptimal. The selection of features must be tailored to the specific market type, with different indicators having varying predictive power in emerging versus developed markets.

**Step 2: Application**

**Main Takeaways of the Results**

1. **Feature Reduction Success**: The most impressive result is that using only 5% of the original features (the Selected(5) subset) achieves better prediction accuracy than using all 216 features. This demonstrates the importance of feature selection in financial prediction models.
2. **Accuracy Improvement**: The Selected(5) feature subset achieved prediction accuracy between 77.82% and 80.27%, which was higher than using the complete set of features.
3. **Computational Efficiency**: When using early stopping techniques with the reduced feature set, the training time was reduced by 84.68% on average while improving accuracy by 13.63%.
4. **Market Similarities**: The Jaccard Distance analysis showed that ETFs from emerging markets (ECH and EWZ) had more similar feature sets (J=0.33) compared to comparisons with the developed market ETF (J=0.64 and J=0.53), suggesting common characteristics across emerging markets.

**Specific Useful Features from the Study**

The most useful features identified across ETFs were:

1. **BBP\_5\_2.0** (Bollinger Band Percent): Appeared in all three ETFs, quantifying price relative to Bollinger Bands
2. **BOP** (Balance of Power): Selected for all ETFs, measures systematic buying/selling pressure
3. **DEC\_1** (Decreasing): Boolean indicator of price decrease compared to previous period
4. **INC\_1** (Increasing): Boolean indicator of price increase compared to previous period
5. **J\_9\_3** (KDJ Indicator): Appears in all ETFs, analyzes and predicts changes in stock trends

For emerging markets specifically:

1. **AOBV\_LR\_2** (Archer's On Balance Volume): Volume indicator for both emerging market ETFs
2. **CTI\_12** (Correlation Trend Indicator): Measures correlation of price with trend line
3. **EBSW\_40\_10** (Even Better SineWave): Measures market cycles using a low-pass filter
4. **K\_9\_3**: Component of the KDJ indicator

For the developed market (IVV):

1. **PVR** (Price Volume Rank): Compares direction of price change to volume change
2. **TTM\_TRND\_6** (TTM Trend): Identifies price bars above/below average
3. **STOCHRSI\_14\_14\_3\_3** (Stochastic RSI): Applies stochastic oscillator to RSI values