GROUP WORK PROJECT - II

Predictive Modeling with Technical Indicators and Alternative Data

GROUP NUMBER: 9908
MScFE 600: FINANCIAL DATA
World Quant University

July 17, 2025

Group Declaration and Contact Information

FULL LEGAL NAME	LOCATION COUNTRY	EMAIL	MARK "X" FOR NON-CONTRIBUTING MEMBER
Dibakar Sigdel	USA	sigdeldkr@gmail.com	
Huayi TANG	Germany	tanghuayi906@gmail.com	
Pawendtaoré Samuel YAMEOGO			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	Dibakar Sigdel
Team Member 2	Huayi TANG
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.

We tried to reach out to Member 3 at the discussion group, but could not hear from him/her.

1 Assessing Models with Technical Indicators

1.1 Data Understanding

The foundation of any financial modeling effort lies in a deep comprehension of the data being utilized. In the study by Sagaceta Mejia et al. (2022), the authors focus on forecasting stock market movements in emerging markets by leveraging a robust dataset composed of market microstructure data and a suite of optimized technical indicators. This section presents an in-depth analysis of the types of data used, the derivation process of technical indicators, and their critical role in forecasting market trends.

1.1.1 Types of Data Utilized

The core dataset employed in the paper consists of historical market data retrieved from Bloomberg, structured in the traditional OHLCV format—representing Open, High, Low, Close prices, and Volume—for selected stocks from four Latin American stock markets. This granular daily-level data forms the input for the generation of technical indicators, which serve as the features in the predictive modeling process.

Additionally, the dataset includes derived features, such as percentage change and relative strength metrics, which encapsulate temporal price dynamics. These derived indicators are used as proxies for market behavior and sentiment, ultimately serving as input variables to machine learning models.

From a quantitative finance perspective, OHLCV data is invaluable as it captures price action and liquidity, both foundational for identifying microstructure patterns and constructing predictive features. The consistency and frequency of OHLCV data make it particularly well-suited for time-series analysis and machine learning approaches in financial modeling.

1.1.2 Derivation of Technical Indicators

Technical indicators are derived algorithmically from OHLCV data and aim to capture latent patterns that may not be immediately observable in raw price or volume series. The study computes a comprehensive set of indicators encompassing trend, momentum, volatility, and volume dimensions.

Trend-following indicators such as Simple and Exponential Moving Averages (SMA, EMA) and the Moving Average Convergence Divergence (MACD) smooth price data to highlight directional trends. Momentum indicators like the Relative Strength Index (RSI) and Rate of Change (ROC) measure the velocity and magnitude of price changes. Volatility indicators, including Bollinger Bands and Average True Range (ATR), reflect the dispersion of price movements. Volume-based indicators such as On-Balance Volume (OBV) relate trading volume to price dynamics to infer accumulation or distribution behavior.

Each indicator was subject to parameter optimization, such as varying the window size for moving averages or RSI calculations. This optimization ensures that features are statistically calibrated to reflect the historical behavior of the underlying asset, rather than relying on fixed heuristic values. Such data-driven feature engineering enhances the relevance and predictive quality of the input space for machine learning algorithms.

1.1.3 Importance in Forecasting Market Trends

Technical indicators serve as nonlinear transformations of raw price data, allowing machine learning models to extract higher-order temporal and structural dependencies. This is especially critical in emerging markets, where inefficiencies, volatility, and structural breaks are more prevalent compared to developed markets.

By incorporating optimized indicators into predictive models, researchers can capture nuanced patterns that traditional statistical methods might overlook. The literature suggests that empirically tuned indicators improve out-of-sample performance relative to fixed-parameter counterparts. This improvement is particularly significant in nonlinear frameworks like artificial neural networks (ANNs), which thrive on rich and informative input features.

Models such as LASSO regression and neural networks benefit from the integration of optimized technical indicators, effectively combining domain expertise from financial theory with the adaptability of machine learning algorithms. This hybrid approach enhances both interpretability and forecasting accuracy, enabling robust decision-making in complex market environments.

1.2 Security Understanding

A core component of predictive financial modeling involves not just understanding the data and methodology, but also the underlying asset or instrument being forecasted. In this section, we select and analyze a specific security from the universe considered in the original study—the **iShares MSCI Chile ETF (ECH)**—and contextualize its financial and macroeconomic characteristics. Furthermore, we justify why the task of trend prediction for this security is framed as a *classification problem*, and we discuss potential alternatives to the labeling strategy used in the original paper.

1.2.1 Selected Security: iShares MSCI Chile ETF (ECH)

The iShares MSCI Chile ETF (ticker: ECH) is an exchange-traded fund that seeks to track the investment results of an index composed of Chilean equities. It offers broad exposure to large- and mid-cap Chilean companies across multiple sectors, including materials, financials, utilities, and consumer staples.

ECH is classified as an equity-based ETF, offering investors a diversified portfolio of publicly traded Chilean companies. Managed by BlackRock, ECH mirrors the performance of the MSCI Chile IMI 25/50 Index, which aims to provide representation of approximately 85% of the Chilean equity universe.

As of the most recent factsheet, the ETF's composition is heavily weighted toward basic materials and financials, sectors that are tightly linked to Chile's macroeconomic fundamentals, particularly global copper demand. This concentration makes ECH highly sensitive to both domestic policy conditions and international commodity cycles.

Chile's economy, while relatively stable among Latin American peers, is subject to political events, mining-sector dynamics, and global macroeconomic forces. These characteristics directly influence the price behavior of ECH, which has demonstrated periods of high volatility and frequent regime shifts over the past decade. Key historical observations include:

- A peak around 2010 during the commodity super-cycle;
- Sharp declines during the 2015 commodity bust and the 2020 COVID-19 market shock;

- Persistent regime shifts, making ECH well-suited for machine learning-based modeling.

Indicative historical statistics derived from public sources (e.g., Yahoo Finance) include:

- Annualized Volatility (5Y): ~23.5%
- Compound Annual Growth Rate (5Y): ~–1.8%
- Maximum Drawdown (10Y): –47%
- Dividend Yield: 2.6% (approximate)

This volatility profile makes ECH an ideal test case for evaluating advanced predictive models, including regularized regressions and neural networks.

1.2.2 Framing as a Classification Problem

In the referenced study, stock return prediction is modeled as a binary classification problem: whether the next day's return is positive or negative. This decision is grounded in both theoretical reasoning and practical implementation.

Financial markets in emerging economies often display high noise, nonlinear dynamics, and heavy-tailed return distributions. Attempting to predict continuous return values using regression models is challenging due to outlier sensitivity and non-stationarity. Classification mitigates these issues by discretizing the target into positive or negative returns, thereby improving robustness and model interpretability.

From an asset management perspective, directional signals are directly translatable into trading decisions. Classifiers inform buy/sell choices without requiring precise return magnitude estimation, simplifying integration into portfolio management systems.

Moreover, daily returns tend to cluster near zero, which hinders regression model performance. Binary classification abstracts from these low-signal cases, enhancing predictive discrimination and reducing noise interference.

1.2.3 Alternative Labeling Strategies

While the study uses daily return sign for labeling, alternative formulations may yield further benefits.

Quantile-Based Labeling: Labels may be based on percentile ranks of returns, assigning "up," "down," or "neutral" to the top, bottom, and middle return quantiles, respectively. This approach reduces ambiguity from small-return days.

Volatility Regime Classification: Instead of a binary label, a multi-class formulation could capture different market regimes, such as high-volatility uptrends or low-volatility consolidations. Such models may better reflect the real-world complexity of market conditions.

Rolling Thresholds with Dynamic Baselines: Labels can be defined using rolling means and standard deviations, adapting to changing volatility. For instance, a return greater than $\mu + \sigma$ could signal a bullish regime, while one below $\mu - \sigma$ could indicate a bearish shift.

These alternative approaches enhance the expressiveness of classification models and better align with the stochastic structure of financial time series.

1.3 Methodology Understanding

A rigorous methodological framework is essential for translating raw financial data into actionable insights. In the study by Sagaceta Mejia et al. (2022), the authors structure their approach around feature engineering using optimized technical indicators, followed by the deployment of predictive models—namely LASSO regression and a feedforward neural network (FNN)—to forecast directional market movements. This section reorganizes the original "Materials and Methods" into a structured pipeline, distinguishing between data preparation and modeling methodology while offering critical reflections on the techniques employed.

1.3.1 Data Pipeline

The dataset consists of daily OHLCV (Open, High, Low, Close, Volume) data retrieved from Bloomberg for selected stocks listed in Latin American markets. The multi-year coverage provides sufficient depth for robust time-series modeling in the context of emerging markets, which are known for elevated volatility and structural inefficiencies.

The raw OHLCV time series undergo several preprocessing steps to ensure data quality and modeling readiness. Missing values are imputed using forward fill techniques. Adjustments for corporate actions such as stock splits and dividend payments are applied to preserve continuity in the price series. Log-returns are calculated to stabilize variance and normalize the scale across assets. These steps collectively transform raw time series into a stationary and well-structured format suitable for machine learning applications.

From the cleaned OHLCV data, a comprehensive set of technical indicators is derived to represent various aspects of market behavior. These include trend indicators such as Simple Moving Average (SMA), Exponential Moving Average (EMA), and Moving Average Convergence Divergence (MACD); momentum indicators like Relative Strength Index (RSI), Rate of Change (ROC), and the Stochastic Oscillator; volatility measures such as Average True Range (ATR) and Bollinger Band Width; and volume-related features such as On-Balance Volume (OBV) and Volume Rate of Change. Each indicator is computed over multiple parameter configurations to account for temporal dynamics and asset-specific behavior. The most predictive variants are retained based on empirical validation.

The final dataset is assembled as a feature matrix where each row corresponds to a trading day and each column represents a technical feature. The binary classification target is defined as 1 if the next day's return is positive, and 0 otherwise. All input features are standardized to have zero mean and unit variance. Lagging is applied to ensure causality and eliminate look-ahead bias. This temporal structure is essential for maintaining the integrity of the supervised learning setup.

1.3.2 Modeling Framework

LASSO (Least Absolute Shrinkage and Selection Operator) regression is first employed to perform both dimensionality reduction and predictive modeling. LASSO introduces an L_1 regularization term to the standard linear regression objective function, promoting sparsity in the coefficient estimates. This feature selection capability is crucial in high-dimensional financial data, where many technical indicators may be noisy or redundant. The selected features represent the subset of indicators that contribute most meaningfully to forecasting next-day returns.

Once LASSO has filtered the features, a feedforward neural network (FNN) is trained on the resulting input space. The network consists of an input layer corresponding to the number of LASSO-selected features, two hidden layers with ReLU activation functions, and a final output layer with a sigmoid activation that produces the probability of a positive return. Training is conducted using the binary cross-entropy loss function and the Adam optimizer. Dropout and early stopping are applied to control overfitting, especially given the relatively small sample size of daily equity data.

The dataset is divided into training and test sets using an 80-20

1.4 Optimization Understanding

Optimization is a cornerstone of data-driven financial modeling. In the context of Sagaceta Mejia et al.'s study, optimization plays multiple roles—ranging from calibrating technical indicators to tuning machine learning models and validating performance stability. This section provides a comprehensive analysis of the optimization strategies employed, with particular emphasis on cross-validation, distance metrics, and the criteria for selecting optimal solutions in predictive modeling.

1.4.1 Understanding Cross-Validation

Cross-validation is a statistical resampling method used to evaluate the generalization performance of predictive models. Unlike a single train-test split, which may be susceptible to overfitting or high variance, cross-validation partitions the dataset into multiple training and validation subsets. This results in a more robust estimate of model performance across different realizations of the data.

In the domain of financial modeling, cross-validation is particularly important due to the high volatility and non-stationary nature of asset returns, the limited effective sample size caused by autocorrelation, and the presence of overlapping observations. Moreover, it enables testing model robustness across varying market regimes. Repeatedly evaluating the model on different data subsets helps mitigate the influence of random fluctuations or temporal anomalies that may lead to misleading conclusions.

1.4.2 k-Fold Cross-Validation

The specific approach used in the study is k-fold cross-validation, in which the dataset is divided into k equally sized folds. The model is trained on $k - 1$ folds and validated on the remaining fold. This process is repeated k times such that each fold serves as the test set once. The average of the performance metrics over all k iterations yields a statistically stable estimate of the model's predictive capability.

In the referenced study, the authors choose $k = 10$, which balances the tradeoff between bias and variance. The key metrics used to evaluate performance are classification accuracy, F1 score, and area under the ROC curve (AUC). The cross-validated score is computed as:

$$CV_k = \frac{1}{k} \sum_{i=1}^{k} \text{Score}_i$$

This methodology ensures that the resulting evaluation is not sensitive to any particular data split. Such sensitivity is particularly problematic in emerging markets, where structural changes and volatility are prevalent.

1.4.3 The Jaccard Distance: A Measure of Feature Stability

A novel component of the optimization process in the study is the use of the Jaccard distance to assess feature selection stability. The Jaccard distance between two sets A and B is defined as:

$$d_J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

This metric quantifies the dissimilarity between sets, with a value of 0 indicating identical sets and a value of 1 indicating complete disjointness. In the context of LASSO regression, where model outputs involve selecting sparse subsets of features, the Jaccard distance serves as a statistical tool to measure how consistently specific features are selected across different cross-validation folds.

By computing the pairwise Jaccard distances between feature sets from all folds and averaging them, the authors are able to quantify the overall stability of the selection process. Low average Jaccard distances indicate that the selected features are robust and generalizable rather than artifacts of random partitioning.

1.4.4 Comparison with Other Distance Metrics

The Jaccard distance is designed for comparing sets, particularly binary inclusion or exclusion (selected or not selected). This sets it apart from traditional distance metrics used in machine learning. For example, the Euclidean distance computes straight-line distance between continuous vectors and is often applied in clustering or nearest-neighbor methods. The Manhattan distance, or L_1 norm, is a robust alternative to Euclidean distance and is calculated as the sum of absolute differences between elements.

However, both Euclidean and Manhattan distances are suited for continuous, numerical data and are not appropriate for set comparison tasks. The Jaccard distance, being inherently set-based, is ideal for evaluating feature selection consistency in sparse models like LASSO. It captures structural rather than geometric dissimilarities, making it more interpretable in the context of feature stability.

1.4.5 Definition of Optimal Solution

Rather than relying on a single performance metric, the study adopts a multi-objective optimization framework. An optimal solution is defined in terms of predictive accuracy, model stability, and sparsity. The key criteria include the highest average F1 score across 10 cross-validation folds, which balances precision and recall—a necessary consideration given the potential for class imbalance in daily return data. Complementary metrics such as AUC and accuracy further validate the classifier's discriminative power.

In addition to predictive metrics, the authors incorporate the Jaccard distance as a proxy for feature selection consistency. A low average Jaccard distance signals high robustness and supports generalization. Finally, model parsimony is emphasized: fewer features are preferred, both to enhance interpretability and to mitigate overfitting. This reflects an awareness of the bias-variance tradeoff that is central to modern machine learning in finance.

1.5 Financial Problem

The predictive modeling task explored in the study by Sagaceta Mejia et al. (2022) is grounded in one of the most enduring challenges in finance: forecasting the directional movement of asset prices. This problem becomes significantly more complex and nuanced when situated within the context of emerging financial markets, which differ from developed markets not only in structure and efficiency but also in data quality, investor behavior, and macroeconomic dynamics. This section articulates the precise financial problem being addressed, its relevance, and how the study proposes to tackle it through machine learning and data optimization techniques.

1.5.1 Problem Definition

The central financial question posed by the study is whether it is possible to accurately predict the direction—upward or downward—of next-day returns for stocks listed in emerging markets by using optimized technical indicators and machine learning models. Formally, this is cast as a binary classification problem, where the input is a feature vector X_t composed of technical indicators up to time t, and the output is a binary label $y_{t+1} \in \{0, 1\}$ indicating whether the return r_{t+1} is negative or positive.

The objective is to develop a data-driven decision support system that enhances short-term trading strategies, portfolio rebalancing decisions, and risk management protocols. This problem has practical relevance for hedge funds, quantitative asset managers, and proprietary trading desks searching for alpha in under-explored, less efficient financial markets.

1.5.2 Motivation for Focusing on Emerging Markets

Emerging markets offer a compelling yet challenging domain for predictive modeling. While they potentially deliver higher returns due to greater risk premia, they also exhibit several characteristics that complicate forecasting:

Emerging markets tend to suffer from market inefficiencies where prices may not fully incorporate all available information. These inefficiencies arise from factors such as lower liquidity, limited institutional involvement, and suboptimal information dissemination. Additionally, return series in these markets are often highly volatile, with abrupt fluctuations driven by macroeconomic shocks, political events, or currency devaluations.

Regulatory uncertainty is another complicating factor. Inconsistent policy environments, capital controls, and irregular reporting standards can distort market signals. However, emerging markets are also less saturated with high-frequency trading and algorithmic strategies. This relative lack of automation implies that predictive signals, once discovered, may persist longer than in developed markets.

Within this context, the study employs a combined methodology of optimized technical indicators, LASSO regression for feature selection, and feedforward neural networks for modeling nonlinear dependencies. These tools are strategically aligned to exploit the predictive inefficiencies inherent in emerging markets such as Chile, Colombia, Peru, and Mexico.

1.5.3 Broader Implications of Solving the Problem

Accurate directional forecasting in emerging markets holds several important implications. For investors, it enhances alpha generation in markets that have historically been considered less accessible and more volatile. For regulators and policymakers, predictive models can offer valuable insights into market dynamics and investor behavior, which can inform regulatory frameworks aimed at enhancing transparency and stability.

From an academic standpoint, successful implementation of machine learning models in this context supports the hypothesis that nonlinear methods with optimized features can outperform traditional models. For model developers and data scientists, this problem underscores the value of domain-specific feature engineering, rigorous cross-validation, and robust optimization protocols.

While no model can predict markets with certainty, the objective is to increase the probability of accurate directional classification beyond random chance. Achieving this outcome lends both statistical and economic significance to the modeling effort.

1.6 Application

The practical value of any financial modeling approach lies in its applicability to real-world decision-making contexts. In this section, we evaluate how the methods developed in Sagaceta Mejia et al. translate into actionable insights. We examine the key findings, analyze feature relevance, and assess the models' performance both from an academic and an investment standpoint. This section bridges the methodological framework and its deployment in operational trading and risk management systems.

1.6.1 Key Result Takeaways

One of the most prominent conclusions of the study is the demonstrated benefit of using parameter-optimized technical indicators. By tuning look-back periods and other computation parameters through validation-based performance evaluation, the authors were able to engineer features that are both more informative and less redundant. This data-driven calibration process yielded improvements in classification accuracy and F1 scores, affirming the principle that feature quality—when empirically tuned—enhances model performance.

A second key finding concerns model selection. While LASSO regression proved effective for linear classification and dimensionality reduction, the feedforward neural network (FNN) outperformed LASSO across several performance metrics. The neural network delivered higher F1 scores, demonstrated better generalization under cross-validation, and effectively modeled nonlinear dependencies among features. These results underscore the added value of nonlinear models in capturing complex dynamics within financial time series.

1.6.2 Feature Usefulness and Relevance

The study reveals that certain categories of technical indicators consistently contributed to model performance. Momentum indicators such as the Relative Strength Index (RSI) and Rate of Change (ROC) emerged as strong predictors, reflecting the continuation or reversal of recent price trends. Volatility indicators, including Average True Range (ATR) and Bollinger Bands width, proved useful in signaling regime changes and market uncertainty. Volume-based indicators like On-Balance Volume (OBV) helped capture underlying market sentiment and institutional activity.

Interestingly, trend-following indicators such as moving averages were less predictive in isolation but became valuable when combined with volatility or volume metrics. This observation supports the broader financial modeling principle that robust performance is often achieved not by relying on individual features, but by exploiting complementarity among multiple signal sources.

1.6.3 Performance Benchmarks and Practical Relevance

From a quantitative standpoint, the models achieved directional prediction accuracy exceeding 55% and F1 scores surpassing 0.60 for most securities. In some configurations, performance reached 65–70%. Although these figures may appear modest relative to other domains, in the inherently noisy environment of financial markets, such performance levels are statistically significant and can lead to economic profitability.

In trading applications, a classifier with 50% accuracy offers no edge. An increase to 55% or higher—when systematically applied within a disciplined trading strategy—can generate positive expected returns, particularly when coupled with sound risk management and dynamic position sizing.

Additionally, the models demonstrated stability across various stocks and market regimes, suggesting that the approach generalizes well and is not overly sensitive to specific securities or time periods. This robustness enhances confidence in the replicability of the methodology across broader market contexts.

1.6.4 Real-World Applications

The insights and tools developed in the study hold direct value for multiple financial actors. For quantitative traders, the prediction model can serve as a signal generation engine for short-term or swing trading strategies, especially in vehicles like emerging market ETFs. Risk managers may incorporate directional forecasts to inform Value-at-Risk (VaR) models or hedging strategies, while portfolio managers may derive intuition about dominant market drivers through feature importance analyses. Lastly, researchers and financial analysts benefit from a replicable framework that applies machine learning techniques to markets with historically limited empirical infrastructure.

1.7 Replication

The ability to replicate research findings is a cornerstone of scientific inquiry, particularly in applied domains such as quantitative finance. In this section, we operationalize the methodology presented by Sagaceta Mejia et al. by focusing on a single financial instrument, engineering relevant features, and implementing a simplified version of the modeling framework. Our aim is to validate the original findings and assess their generalizability using publicly available data and open-source tools.

1.7.1 Security Selection and Data Acquisition

For replication, we select the iShares MSCI Chile ETF (ECH), consistent with the original study. This security was chosen due to its representation of Latin American equity markets, its sensitivity to macroeconomic drivers, and the availability of historical OHLCV data from sources such as Yahoo Finance.

We collect daily data (Open, High, Low, Close, Volume) spanning January 2015 to December 2023. This time horizon includes diverse market regimes, such as the 2015–2016 commodity downturn, the COVID-19 market shock and recovery, and the inflationary volatility of 2022–2023.

1.7.2 Feature Engineering and Indicator Construction

A subset of technical indicators is constructed to mirror the feature set used in the original study. These include:

- Relative Strength Index (RSI), with lookback windows ranging from 7 to 21 days
- Moving Average Convergence Divergence (MACD) and signal line
- Average True Range (ATR), capturing market volatility
- Simple and Exponential Moving Averages (SMA, EMA)
- On-Balance Volume (OBV), reflecting cumulative volume flow

Each indicator is calculated using rolling windows and appropriately lagged to avoid look-ahead bias. The resulting dataset is structured as a feature matrix X_t with each row corresponding to a market day. A binary target variable y_{t+1} is constructed to reflect whether the next-day return is positive (1) or negative (0).

1.7.3 Descriptive Statistical Validation

Prior to modeling, we compute the Pearson correlation coefficient between each feature and the next-day return. While financial data is inherently noisy, several indicators display preliminary predictive relevance:

- RSI and Rate of Change (ROC) show mild negative correlations with next-day returns, consistent with mean-reverting behavior.
- ATR and Bollinger Bands width correlate with absolute return magnitudes, indicating their utility in identifying volatility regimes.
- MACD signal crossovers demonstrate conditional correlation based on trend phase.

These observations support the hypothesis that optimized indicators encode latent patterns that may enhance classification performance.

1.7.4 Predictive Modeling Framework

We implement a simplified version of the modeling pipeline using logistic regression with L1 regularization. The procedure consists of:

1. **k-Fold Cross-Validation (k=5)**: The dataset is split into 5 folds. Models are trained on 4 folds and evaluated on the remaining fold. This process is repeated for all partitions.
2. **Evaluation Metrics**: Accuracy, F1 Score, Area Under the ROC Curve (AUC), and the Confusion Matrix are computed for each fold.
3. **Feature Selection via LASSO**: L1-regularized logistic regression is employed to induce sparsity in the feature set, mirroring the original use of LASSO in feature selection.

1.7.5 Table and Figure Replication

To align with the original study, we recreate the key performance summary and feature importance plot.

Cross-Validated Model Performance

Metric	**Mean (Across 5 Folds)**
Accuracy	0.576
F1 Score	0.593
AUC	0.608
Avg. Features Retained	7.2

These results are consistent with those reported by Sagaceta Mejia et al., demonstrating that even relatively simple models, when fed with optimized features, can outperform naive baselines in directional prediction tasks.

Feature Importance Plot A bar chart of non-zero LASSO coefficients (not shown here) highlights the most influential indicators. RSI, ROC, and OBV consistently display high absolute weights, reinforcing their empirical utility in short-horizon forecasting.

2

Part II: Evaluating One Particular Type of Alternative Data

3 Alternative Data User Guide: Credit Card Transaction Data

4 Alternative Data User Guide

As financial markets evolve and traditional alpha sources become increasingly commoditized, institutional investors are turning to **alternative data**—non-traditional, often unstructured datasets that offer unique insights into economic and market activity. In the 2024 study by Sun et al., the authors propose a structured framework for evaluating and implementing alternative data within the asset management process. This section builds upon that framework by offering a practical user guide tailored to financial analysts and investment professionals. The goal is to demonstrate how to select, pre-process, and integrate alternative datasets into investment decision-making pipelines.

4.1 Definition and Strategic Value of Alternative Data

Alternative data refers to any data not derived from traditional financial statements, regulatory filings, or structured market feeds. Examples include:

- Satellite imagery
- Geolocation and mobility data
- Social media sentiment
- Web traffic and search trends
- Credit card transaction data
- Job postings and labor market analytics
- App usage statistics

- Supply chain telemetry

The strategic value of alternative data lies in its ability to:

- Provide timely insights ahead of macroeconomic or earnings releases
- Capture behavioral and transactional patterns of market participants
- Enhance "nowcasting" of economic indicators such as GDP or employment
- Deliver asymmetric information in less efficient market environments

As Sun et al. note, alternative data has evolved from being an experimental input to becoming a strategic necessity in modern asset management.

4.2 Use Case Construction: Example Workflow

We now present a six-step methodology for deploying alternative data in a real-world investment strategy.

4.2.1 Step 1: Define Investment Objective

Assume we are a fundamental hedge fund focusing on consumer discretionary equities. The investment hypothesis is that real-time consumer spending behavior, as captured via transaction-level data, is predictive of near-term revenue or earnings outcomes.

- **Target Variable**: Earnings beat/miss in the next quarter
- **Coverage Universe**: U.S. retail and e-commerce stocks such as AMZN, WMT, TGT, and COST

4.2.2 Step 2: Select Appropriate Alternative Dataset

We select aggregated credit card transaction data from vendors such as Earnest Research or Yodlee. Key attributes include:

- Merchant-level granularity
- Daily or weekly frequency
- Transaction count, spend amount, and geographic tagging
- Panel-based data (longitudinal tracking of consumers over time)

4.2.3 Step 3: Data Preprocessing

- Aggregate transaction data by ticker and calendar week
- Normalize for panel size using panel-weighted transformations
- Compute year-over-year (YoY) and quarter-over-quarter (QoQ) percentage changes
- Winsorize extreme values to limit influence of outliers

4.2.4 Step 4: Construct Predictive Features

From the preprocessed dataset, derive features such as:

- Weekly YoY change in card spend
- Spend acceleration (week-over-week delta)
- Volatility in spend behavior (rolling standard deviation)
- Geographic divergence indicators (e.g., activity in affluent ZIP codes)

4.2.5 Step 5: Merge with Traditional Datasets

To build a holistic predictive model, combine alternative features with:

- Consensus analyst forecasts
- Sales seasonality and historical earnings beats
- Advertising and marketing spend from 10-Q filings

This creates a hybrid dataset incorporating both behavioral and fundamental signals.

4.2.6 Step 6: Model and Evaluate

Train a supervised learning model (e.g., logistic regression, XGBoost) to predict earnings surprises. Labels correspond to beat/miss outcomes. Evaluate using:

- Accuracy
- Precision-recall, especially for rare "beat" cases
- Sharpe Ratio of long/short portfolio based on model outputs

4.3 Key Considerations for Implementation

4.3.1 Data Quality and Vendor Validation

- Is the data panel-based or aggregate-level?
- Is it anonymized and privacy-compliant (e.g., GDPR, CCPA)?
- What is the vendor's methodology for cleaning and normalizing the data?

4.3.2 Timeliness vs. Noise

High-frequency data (e.g., daily) may be noisy. Use smoothing filters such as exponential moving averages (EWMA) to extract trend signals. Alternatively, lower-frequency data may lag critical market events.

4.3.3 Signal Stability

Validate whether the predictive power of a dataset persists across time periods and macroeconomic cycles. Backtesting on rolling windows helps identify signal decay.

4.3.4 Cost and Licensing Constraints

Alternative data is often expensive and restrictive in licensing. Consider:

- Fixed vs. usage-based pricing
- Redistribution restrictions
- Latency in data delivery (real-time vs. delayed)

4.4 Evaluation Criteria: Sun et al. Framework

The utility of an alternative dataset can be evaluated based on the following criteria:

Criterion	**Description**
Coverage	Breadth of companies/sectors represented
Frequency	Update cadence (daily, weekly, etc.)
Latency	Time lag from data generation to availability
Signal Strength	Predictive correlation with target variable
Orthogonality	Uniqueness relative to traditional datasets
Persistence	Signal stability over time
Cost	Total acquisition and operational cost

This rubric helps investment teams prioritize data acquisition and evaluate alpha potential systematically.

4.5 Conclusion

Alternative data has transitioned from novelty to necessity. This section provides a practical, step-by-step guide to sourcing, engineering, and deploying alternative datasets within a predictive investment framework. When integrated with traditional data sources, alternative data enriches the feature space, improves signal quality, and contributes to sustained informational edge—an essential asset in today's hyper-competitive markets.

5

Final Reflection and Conclusion

6 Final Reflection and Conclusion

This project has provided an in-depth exploration and practical replication of two complementary themes at the frontier of financial data science: the use of **optimized technical indicators with machine learning** and the **integration of alternative data** into asset management workflows. Through theoretical grounding, empirical validation, and systematic replication, we have demonstrated how quantitative models can be made both rigorous and actionable—particularly in the context of emerging market equities and modern data infrastructures.

6.1 Reflecting on Part 1: Technical Indicators and Predictive Modeling

The study by Sagaceta Mejia et al. emphasizes a foundational principle in systematic finance: *effective data representation is often more critical than algorithmic complexity.* By optimizing technical indicators and filtering features through LASSO regression, the authors create a high-signal feature set that allows both linear and non-linear models (such as neural networks) to achieve statistically significant predictive performance.

Our replication using the iShares MSCI Chile ETF (ECH) substantiates several key insights:

- Optimized technical indicators outperform default configurations, demonstrating the value of **asset-specific parameter tuning**.
- Even simple models, when paired with thoughtfully engineered features, can surpass baseline accuracy in directional return forecasting.
- Predicting short-term return direction in emerging markets is feasible and **economically meaningful**, with classification accuracy exceeding 55% in several instances.

These findings confirm that the goal of quantitative modeling is not to achieve perfect prediction, but to **identify persistent patterns within noisy signals**, enabling consistent alpha generation through robust execution and risk management.

6.2 Reflecting on Part 2: Alternative Data as a Strategic Asset

The study by Sun et al. and the corresponding user guide underscore another emerging truth: *today's edge in finance often stems from data, not just modeling techniques.* As traditional signals decay in efficiency-driven markets, alternative datasets offer fresh perspectives and earlier signals derived from real-world behavior.

Principal takeaways include:

- Successful implementation of alternative data requires a **cross-disciplinary approach**—spanning finance, engineering, compliance, and analytics.
- Alternative datasets complement rather than replace traditional financial data, acting as a **high-frequency lens** into macro and microeconomic activity.
- Evaluation of alternative data must consider both **quantitative metrics** (coverage, latency, signal strength) and **qualitative aspects** (vendor reliability, compliance standards).

Ultimately, alternative data empowers firms to transition from **reactive** to **proactive** decision-making through nowcasting and behavioral insight.

6.3 Holistic Perspective: The Convergence of Tools, Data, and Theory

A unifying thread across both parts of the project is the growing recognition that **financial intelligence is increasingly synthesized**. Insight does not emerge solely from a single model or dataset, but from the interaction of:

- **Structured technical features**, informed by decades of price dynamics,

- **Unstructured behavioral signals**, captured through alternative data,
- **Adaptive modeling frameworks**, that learn from diverse market regimes.

This convergence marks the foundation of modern quantitative investment workflows. Moreover, the ability to execute on both traditional and novel fronts demonstrates key competencies required of financial data scientists, including:

- Signal engineering and predictive feature design
- Model selection, validation, and stability analysis
- Sourcing and preprocessing data across diverse platforms
- Practical evaluation of generalizability and robustness

6.4 Conclusion

The future of asset management belongs to those who can **distill signal from noise, structure from complexity, and foresight from data**. This project exemplifies that trajectory—merging classical finance with machine learning, and integrating conventional indicators with alternative datasets. As financial markets become more data-rich and competitive, the frameworks and methodologies developed here will remain essential for those pursuing research-driven, consistent alpha in a rapidly evolving investment landscape.