Feature Selection in Finance, It is Delicious

Benjamin A. Schifman
Justin J. Siekmann
Department of Electrical and Computer Engineering
University of Arizona
Tucson, AZ 85719
bschifman@email.arizona.edu
jsiekmannemail.arizona.edu

Abstract

The abstract paragraph should be indented ½ inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1 Introduction

There exist technical analysis indicators traditionally used by analysts to evaluate and predict market and equity performance. These indicators provide a unique perspective on the strength and direction of the underlying price action in market data. Feature selection is used to determine relevant indicators while identifying those that are irrelevant and redundant. Different implementations of algorithms based on these indicators could be used to predict performance of individual equities, sectors, or overall markets. They could also be used to classify and identify the correlation and interdependencies between equities, sectors, and markets. The goal of this paper is to implement various approaches to determine the efficacy of technical indicators as enablers to financial analysis. For analyzing features This project presented a couple challenges during implementation including developing an accurate testing method as well as handling and computing such large volumes of data.

Possibly reword and keep/move:
From this project we hope to deepen our understanding of the usage cases for applying specific machine learning algorithms as well as expanding upon our technical analysis of the stock market and which indicators play a role in successful market analysis.

2 Related Work

This is optional. If wanted, save til last.

3 Methods/Approach

The following subsections present details and explanations of the methods and functions implemented as part of this project.

3.1 Data and Technical Analysis Indicators

The Quandl platform was used to fetch 11 years of market data in total from Dec 31, 2006 to March 27, 2018 on various identified US tickers across different sectors. Tickers used in the project can be found in Table 1 categorized by sector. As the process to fetch and preprocess the data is time

ECE523: Engineering Applications of Machine Learning and Data Analytics (Sp2018).

consuming, pickle files were used to save data locally to be quickly reimported. TA-Lib: Technical Analysis Library was used to calculate features on the market data for each ticker. TA-Lib has the ability to calculate many technical analysis indicators in various categories. The features incorporated into this project, found with in Table 2, are a selected subset of the indicators offered in TA-lib based upon the categories in TA-Lib, popularity online, and expert's favorites and essentials.

Financials	Utilities	Energy	Healthcare	Technology	Real Estate
JPM	T	XOM	JNJ	AAPL	ECL
BAC	VZ	CVX	UNH	GOOGL	FMC
WFC	NEE	BP	PFE	MSFT	IP
C	TMUS	GE	MRK	FB	VMC
MS		SLB	ABBV	INTC	BMS
			MMM	CSCO	
			AMGN	ORCL	
			MDT	IBM	
				NVDA	

Table 1: Tickers

Overlap	Momentum	Volume	Cycle	Price	Volatility	Statistical	Math Transformation	Pattern Recognition
SMA	RSI	OBV	HT_DCPERIOD	AVGPRICE	ATR	BETA	EXP	CDLENGULFING
EMA	MOM	AD	HT_TRENDMODE	TYPPRICE		LINEARREG	LN	CDLDOJI
BBAND_Upper	ROC	ADOSC				VAR		CDLHAMMER
BBAND_Middle	ROCP							CDLHANGINGMAN
BBAND_Lower	STOCH_SLOWK							
HT_TRENDLINE	STOCH_SLOWD							
SAR	MACD							
	MACDSIGNAL							
	MACDHIST							

Table 2: Technical Analysis Indicators

3.1.1 Indicator Categories

Overlap indicators generally can be overlayed onto the price charts. They most commonly include different styles of moving average calculations.
Momentum indicators convey how quickly the price of the ticker is moving. For exampe, the faster the price of a ticker increases the larger its momentum.
Volume indicators take into account the volume of the day's trading into account.
Cycle indicators attempt to identify changes in the overall direction of the ticker's movement.
Price indicators combine the multiple prices in the data into one value.
Volatility indicators convey how sporadic the ticker's prices are.
Statistical indicators are based on statistical concepts and can be used in a number of different ways.
Math Transformation indicators apply common mathematical operations upon the tickers prices.
Pattern Recognition indicators look for patterns in the prices of a ticker traders have identified are indicative of future outcomes.

3.2 Normalization

As each technical analysis indicator produces values applicable based on how the indicator was calculated, normalization of the indicators makes correlations between them during feature selection more accurate and applicable. Each value in a specified time period is normalized using Equation (1). If there is extra data not consisting of a full time period, the extras are thrown out at the beginning of the data as data near the end may be more relevant and thus more desirable to keep. The start indices are computed for each ticker and return to ease future handling.

$$x_n = \frac{x - min}{max - min} \tag{1}$$

3.3 Feature Correlation

Removing highly correlated features allows for the optimization of the classification algorithms by reducing the feature space. Features that are highly correlated most likely offer no additional data and

they are an extra expense in computation time. The pairwise correlation of columns was computed, and columns that were correlated above a certain threshold were marked to be removed from future classification algorithms. The following Figure 1

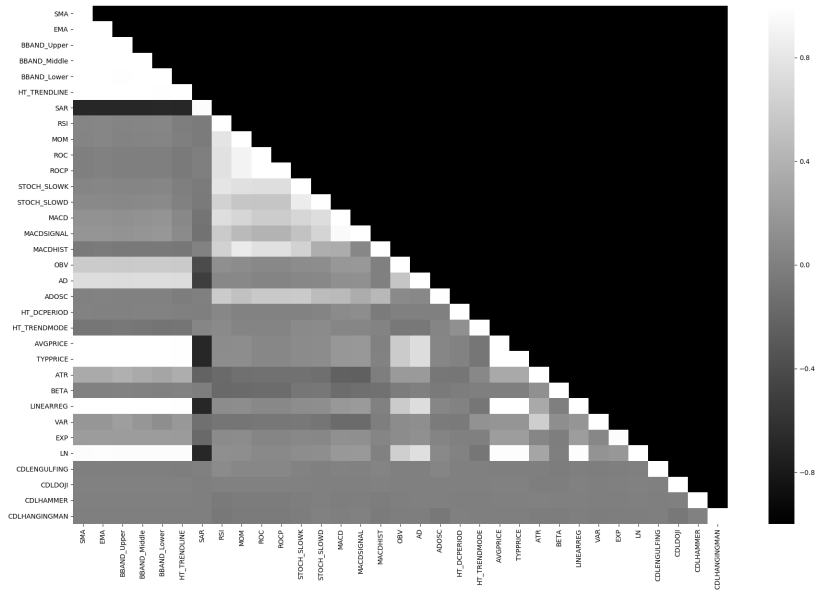

Figure 1: Heat Map of Correlated Variables

3.4 Maximal Information Coefficient (MIC)

The MIC is a measure of two-variable dependence designed specifically for rapid exploration of many-dimensional data sets [1]. A benefit to MIC correlations between two variables is that it can be described regardless of linear or non-linear relationships. The MIC yields a single vale $0 \leq MIC \leq 1$ with a value closer to 1 representing that the variables are more closely correlated, and a value near 0 indicates statistically independent variables that have neither linear nor nonlinear relationships. The $minepy$ library was used in python to rank the features according to their MIC with the target variable. The MIC was calculated for each feature in each ticker, yielding 33 MIC values for each ticker, and then a final set of MIC values were calculated by taking the mean of all the sets of MIC values over all of the tickers.

3.5 Recursive Feature Elimination (RFE)

RFE is a method of recursively selecting smaller subsets of the larger feature set to then be used in an external classifier. The goal of RFE is to reduce the feature set into the smallest set of relevant and valuable features which yield the greatest accuracy. The time to run of the external classifier is reduced due to the fact that there are fewer features analyzed in the classification.

3.6 Random Forest Classifier (RFC)

The RFC is an ensemble algorithm implementing a combination of decision trees classifiers where the majority vote of all trees are used to classify the input feature vector [2]. The RFC was used in order classify the input data, while also intrinsically implementing feature selection to return a vector of feature importances as depicted in Figure 3.

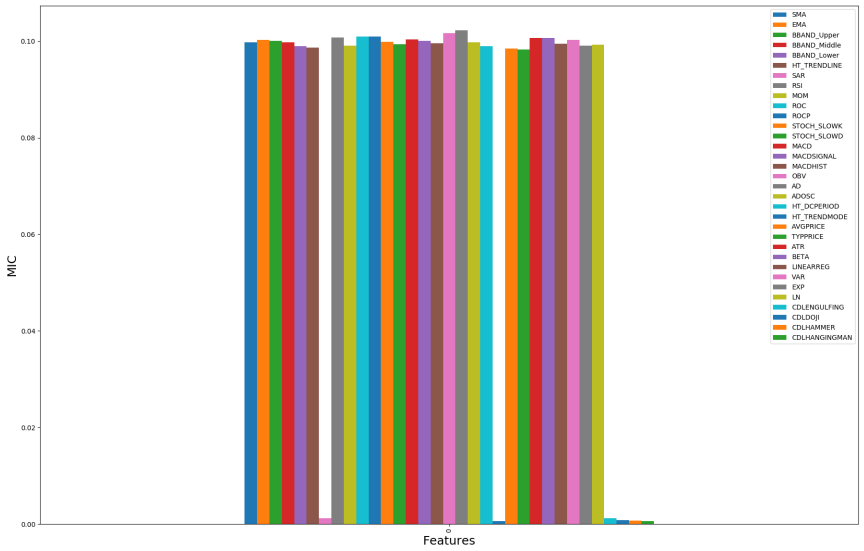

Figure 2: Maximal Information Coefficient

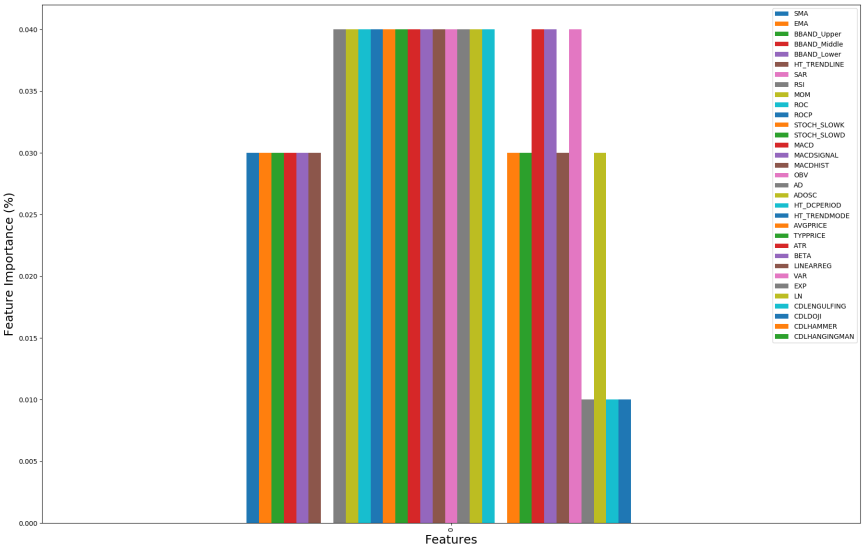

Figure 3: Random Forest Classifier

3.7 Principle Component Analysis (PCA)

PCA orthogonally transforms a set of features into a set of linearly uncorrelated principal components. PCA is a method for reducing the dimensionality of the feature set size while retaining principal component variance, and the features informational relevance. To analyze the performance of this method, PCA was implemented on the original 33 features and then the resulting principal

components were used in the RFC classifier with 5 fold cross validation to see the resultant accuracy. The number of principal components were iterated from 1 to 33 and then the principal components were implemented into the RFC to find the optimal number of principal components to yield the best accuracy and in the least amount of time. The following Figure 4 depicts the PCA accuracy and time taken with the RFC vs. the number of principal components.

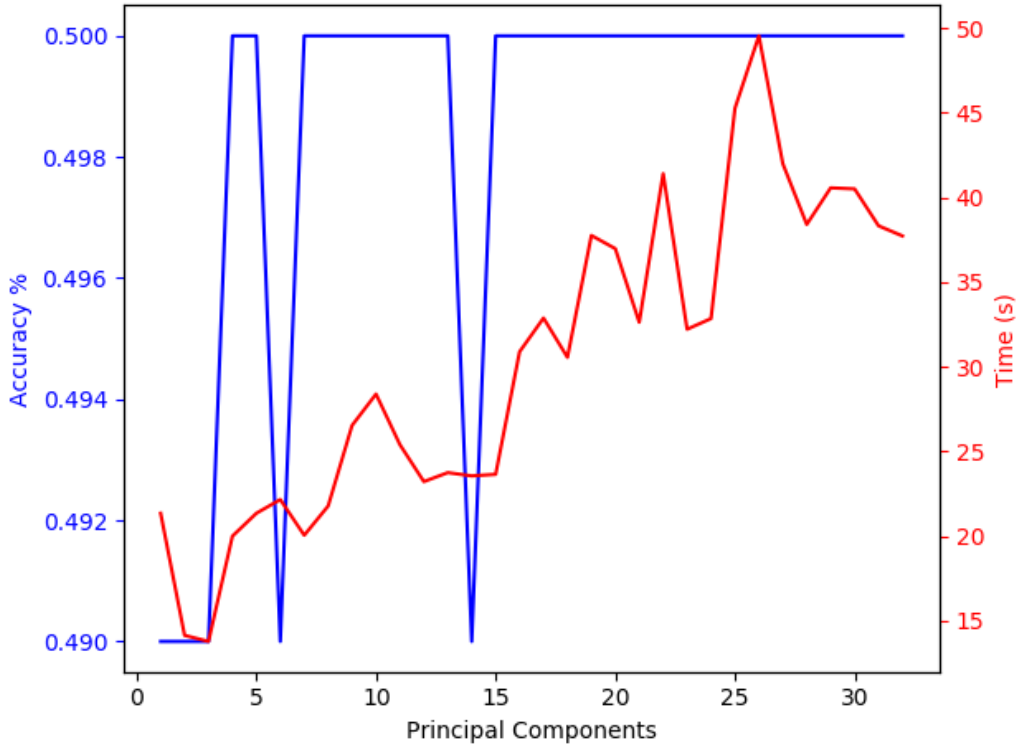

Figure 4: Principal Component Analysis

4 Results

5 Conclusion

References

References

[1] David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. Detecting novel associations in large data sets. *science*, 334(6062):1518–1524, 2011.

[2] Mahesh Pal. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1):217–222, 2005.