Analyzing the Predictability of Exchange Traded Funds Characteristics in the Mutual Fund Market on the Flow of Shares using a Data Mining Approach

Asil Oztekin  
University of Massachusetts Lowell  
asil_oztekin@uml.edu  

Kyle Best  
University of Massachusetts Lowell  
Kyle_Best@uml.edu  

Dursun Delen  
Oklahoma State University  
dursun.delen@okstate.edu

Abstract

This study is aimed at determining the future share net inflows and outflows by using the characteristics of Exchange Traded Funds (ETF) as variables in a data mining based analytic methodology. The relationship between net flows is closely related to investor perception of the future and past performance of mutual funds. In order to explore the relationship between investor’s perception of ETFs and subsequent net flows, this study is designed to shed light on the multifaceted linkages between fund characteristics and net flows. An international selection of 222 ETFs from one of the top three ETF providers is used in this study, of which fifteen attributes from each fund are used because they are likely to be contributors to fund inflows and outflows. Cross-Industry Standard Process for Data Mining (CRISP-DM) is used in this study accompanied with machine learning tools to develop a neural network which will forecast a positive or negative flow of net assets for ETFs.

1. Introduction

In recent years the competition between mutual fund providers and those entering the ETF sub market has become increasingly competitive [1]. Competition can most clearly be seen in the expense battle between large financial institutions such as Vanguard and BlackRock. The underlying reason for reducing the expenses of the funds is to attract investors. The structure of an Exchange Traded Fund is slightly different from your traditional mutual fund because its main objective is to meet the performance of an underlying index. There has been little research into why an investor chooses one ETF over another, especially those that have similar holdings and track like indices. Knowing the characteristics of a fund that are most important to investors can be used in marketing techniques and the development of new funds that highlight favorable characteristics. Having the right combination of key fund characteristics is expected to bring inflow of shares and increase the total net assets of the fund. For a financial institution developing and marketing these financial products, the goal is to increase net assets year over year. If a fund’s total net assets increase over time, they are deemed to be profitable investments. For those funds that lack to attract positive share inflows financial institutions backing such products are faced with closing the funds and thus never realizing a return on their investments.

Most academic research has been focused on determining the flows of mutual funds through past fund performance based on returns earned. In the studies reviewed by Nanigian [2] fund performance has been explained to have a linear relationship with fund expenses with an increase in expenses having a negative impact on fund performance. Yet, Nanigian [2] believes these recent studies in regression analysis are assuming a linear relationship that in actuality is far from linear and more complex. This supports the need for a method of machine learning that can discover relationships that are not linear in nature. Cooper et al performed a study on the effects of recent name changes on the net flows of a fund. In this study Cooper et al found results through the use of regression techniques, indicating funds experienced an abnormal inflow of funds just after a recent name change proving there is more to fund flows than just performance [3].

ETFs are developed to meet the performance of an underlying index and not necessarily outperform the market, and thus their attractiveness to investors is not of mere performance of the fund but its broader set of characteristics. Data mining approaches can be of significant use in determining why a particular ETF attracts more investors than others by using machine learning to bring to the forefront linkages and correlations of fund characteristics that have yet to be documented. Therefore, this study is aimed at developing a data mining-based methodology to determine the prevailing factors of an attractive ETF.

The rest of the paper is structured as follows. Section 2 briefly summarizes the recent and relevant
literature in the field. Section 3 describes the main steps in our proposed data mining methodology. Section 4 presents the modeling results and discusses the findings. The paper concludes with Section 5 where a summary of the study and its implications are discussed.

2. Literature Review

Past studies in academia have focused their efforts on fund flows as a result of one or two independent variables. These studies have used linear regression models to explain the relationship between variables and fund flows. When multiple variables were considered in a study they were conducted separately and not together in their analysis suggesting that each is independent of one another with no influence of one variable on the relationship of another. Sirri et al. [4] conducted a study over a 30 year period to determine the relationship between fund performance and fund flows. They also looked at the relationship between fund flows and fund’s provider as a source of influence on flows. The results indicated that past fund performance has a positive relationship on fund flows. It also indicated fund provider membership in well-advertised campaigns increased fund flows. Fund performance as a determinant in fund flows has been extensively studied by academics and its positive relationship is not only supported in Sirri et al. studies but of many others. Additional research done by Cashman et al. [5] has similar results to Sirri et al. The Cashman et al. study [5] shows how investors rely on performance as a determining factor for a buy or sell decision. A study conducted by Cha et al reports much of the same results. They concluded in the overall US market, investors tend to put their money into funds based on past security performance [6].

Cici [7] conducted a study on US equity mutual funds to see if there is a tendency for investors to sell their mutual fund holding in times of positive performance and hold onto the funds during times of negative performance. Cici [7] called this the disposition effect in which investors tend to hold onto losses and realize gains. Cici [7] compares findings with that of Grinblatt et al. [8] using logit regression and concludes the same with past performance being a contributing factor in the flow of shares. Supporting Sirri et al is field work done by Hoffmann et al [9]. Their study sheds new light on the psychological factors influencing the prevalence of mutual funds flows and fund performance. The study reveals investors commonly perceive previous positive fund performance will continue into the future and thus they hold on to mutual funds with past strong performance and dump those with negative or weak past performance. This helps to explain why past fund performance has a positive relationship on fund flows.

Gottesman et al. [10] explore the relationship between investor behavior and fund flows in up and down markets. Gottesman et al. study [10] actively managed mutual funds and the net flow of funds in and out of the funds during economic downturn and economic upturns. The conclusion was investor behavior influences the net flow of funds disproportionately in up versus down markets. In down markets investors are less likely to contribute to inflows even if the fund out performed its peers than if it were an up market. Again this supports the work of Hoffmann et al indicating investor behavior is a key attribute in predicting future fund flows and needs to be considered in the development of a more comprehensive study on fund characteristics.

Daily fund flow volatility and its effect on fund performance were explored by Rakowski [11]. Rakowski [11] conducted a study on US open-ended mutual funds to see if there is a correlation between daily fund flow volatility and performance [10]. Weak performance was due to the increase in trading and expenses incurred as a result of fund volatility. Huang et al also studied fund volatility. Huang et al argued that fund volatility affects fund inflows negatively as investors view fund volatility as a detriment to predicting future fund performance [12]. Bollen [13] takes a targeted approach to fund flows by segmenting the US mutual fund market and focuses on the socially responsible mutual fund niche. Bollen [13] uses a flow performance regression technique to assess the volatility of fund flows for a sample of socially responsible mutual funds. The results indicate that socially responsible funds are more resistant to the fund flow volatility than conventional funds. These results indicate that the type of fund and its meaning to investors has a correlation to overall fund flows. Therefore one can extrapolate there are more attributes to a fund that affect the fund flows than that of just past performance. It brings to the surface the need to do further research on the subject of fund flows and to broaden the variables tested. One of the best ways to determine variables effecting fund flows aside from past performance is through the use of funds that do not exist to outperform the market. ETFs are such funds designed in this way and thus are an excellent source to test.

A slightly different approach to the influences on fund flows is conducted in a study by Kempf et al. [14]. The study takes into consideration the family of funds a mutual fund resides in as a predicting factor
in fund flows. It also speculates that the position within the fund family also has influence on fund inflows. In order to create a model which tests only the variable of fund family and position Kempf et al. [14] attempt to control fund characteristics such as expense ratios, fund size, and age. While it creates a more focused test it does dispel the need to test fund characteristics as determining factors in fund flows. As a result, Kempf et al’s research only partially explains fund flows because they build their model purposely to leave out these factors.

Interesting research on the topic of fund characteristics to fund performance was conducted by Fan et al. The study took fund characteristics such as expense ratios, management tenure, and fund age and compared them to the resulting fund performance [15]. Popular beliefs would lead investors to believe these three fund characteristics would have an impact on fund performance but the study findings suggested they were irrelevant. Berk et al conducted a study on mutual fund flows and performance in rational markets [16]. Of specific concern was the ability of active mutual fund managers to outperform the market. Their test results support the use of actively managed funds by superior fund managers and the tendency of investors to put their money in funds that outperform their peers. Yet, Berk et al. strived to explain the investor behavior as rational versus irrational behavior. This indicates investors are actively researching and making investment choices that affect fund flows based of fund characteristics such as past performance.

How broker incentives affect inflows is a question addressed by Christoffersen et al. [17] in a study conducted to determine if the incentives of brokers truly influenced the inflows of affiliated funds. Logic would expect the results of such a test to reveal that incentives do influence fund flows. Christoffersen et al. [17] used time series regression to test their hypothesis. They concluded that flows to affiliated brokers are less sensitive to the fees and or commissions of the particular affiliated fund compared to unaffiliated counterparts. These results indicate that broker incentives are an attribute to be considered in predicting future fund flows with brokers receiving above average incentives for a fund which in turn directly influences fund flows.

Barber et al. cross section regression shows investors do make investment decisions based on different types of fund expenses [18]. In general investors tend to favor mutual funds with lower commissions and up-front fees but they do so because of how the fees are presented to the investors. Commissions are more readily identifiable by investors and such are more influential in the decision on mutual fund selection creating more inflows to funds with lower commissions than alternatives with higher commissions.

The majority of these studies concluded that past performance is a key contributor in predicting future fund flows as investors tend to trade based off past performance. While this is true for traditional mutual funds ETF’s are a different animal and past performance is expected to be less of an influence because the objective of the ETF funds is to not outperform the market. As a result the recent research done on volatility, expenses, fund family, and type of underlying securities of the mutual fund appear to be fund characteristics that influence investor behavior and the ultimate net flow in or out of a fund. Therefore this study is aimed at performing a more comprehensive review of fund characteristics and how they relate to net fund flows of ETFs.

3. Methodology

This study uses the well-known CRISP-DM methodology in conducting research and reporting results. The Cross-Industry Standard Process for Data Mining is used as the method to conduct research because of its acceptance in the industry as a standard for data mining with the backing of industry leaders. CRISP-DM was developed with the help of a special interest group with a common interest to create a methodology that could be used as a framework for conducting data mining projects [19]. CRISP-DM is designed as a complete framework for data mining projects which is the reason for its use in this study. It not only provides the means to conduct the study but also provides assurance of a professional design that has been tested by leaders in the industry. The methodology calls for the use of a six step process. Each step builds upon the previous creating a well thought out process design and implementation. The six steps include business understanding, data understanding, data preparation, model building, testing and evaluation, and deployment. Each step is performed in consecutive order with the option of circling back to previous steps if the business problem calls for additional clarity or adjustments in the data. Figure 1 depicts a graphical representation of the six step CRISP-DM method. The arrows indicate that each step builds upon the previous and the order of the data mining steps. The nature of the CRISP-DM model is based on a complete analysis at each step before continuing on to the next to avoid any flaws in the data mining project that could lead to inaccurate findings.
Step 1: The first step of the CRISP-DM process emphasizes the need to have a complete understanding of the business in which the data mining is conducted. Objectives, goals, and a plan are determined. In this study the general topic of research is the net inflows and outflows of mutual funds and is lesser known counterpart exchange traded funds. The objective is to determine the characteristics of the fund that influence the buying and selling decisions of investors. The ultimate goal of this data mining project is to use the results of data mining methods to modify and create new exchange traded funds that attract net inflow of funds from investors.

Step 2: In the second step of the CRISP-DM process data is collected and verified for accuracy and consistency. In this study data is collected from confidential sources and compiled based off its expected correlation between fund flows and their characteristics. Much of the data selected for the study is based on the professional experience with data analysis of exchange traded funds and through a thorough understanding of ETF’s as well as their differences as compared to traditional mutual funds.

Step 3: Data preparation is then conducted in the third step of the CRISP-DM process. Data to be used in the study is selected, cleaned, and formatted for modeling. A detailed explanation of the data formatting and extraction is discussed in the forthcoming section 2.1.

Step 4: After data has been organized and formatted the next step is to select and build the models for the study. The models selected in this study are used because of their ability to analyze large sets of variables in literature. As previously mentioned on the of reasons for conducting the study is to use a data mining approach that can incorporate a large set of variables that previous studies were lacking. Decision trees, artificial neural networks from the data mining area are actually the most commonly used methods due to their high prediction accuracy, efficiency in computation, and relatively easier explanation.

Step 5: Step five includes the evaluation of results from the various models. Review the overall process for adequacy and to determine the next steps. Included in the results and findings is sensitivity analysis which is critical in answering one of the main objectives of this studying by identifying the variables that influence the results of the data mining methods used. If you recall an underlying assumption is that Exchange Traded Funds net flows are dependent on the characteristics of the fund. Investors view characteristics as being a positive or negative in respect to their investment strategy or portfolio.

Step 6: The final and six step of the CRISP-DM process is to report the final results of the project in a way that can be used in decision making processes at the management level. In particular, this study’s results will be used in selecting characteristics which will be used to create new products to be developed and marketed to investors.

3.1. Data Understanding and Preparation

In steps two and three of the CRISP-DM process data is selected and formatted for use in data mining models. Much of the data is extracted from confidential sources while the rest is taken from publicly available information from iShares.com and Yahoo Finance. Each variable is considered with the overall goal of the study in mind. The variables should have influence over the decision of investors in choosing where and which financial product they wish to place their investment dollars. Therefore the reason for variable selection is based off logical inferences. Table 1 lists the variables used in this study along with the variable type and short explanation of each.

The reason each variable is used is now fully explained starting from the top of Table 1. The fund type variable is used because it is a broad representation of the underlying securities in the fund. The fund type indicates whether a fund’s majority holdings consist of US securities or
foreign/global securities. It also indicates whether the securities are fixed income or equities. Investors have different reasons for investing in each type of security. Often the case for fixed income securities is to have a steady stream of income in contrast to equity holdings that re-invest more of their earnings into future growth in which the investor is looking for long term price appreciation.

Table 1. Variables Used in this Study with Type and Explanation

<table>
<thead>
<tr>
<th>Variable</th>
<th>Variable Type</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fund Type</td>
<td>Nominal</td>
<td>Type of underlying securities</td>
</tr>
<tr>
<td>Family of Index Being</td>
<td>Nominal</td>
<td>Underlying index family</td>
</tr>
<tr>
<td>Tracked</td>
<td>Numeric</td>
<td>Age of fund since inception</td>
</tr>
<tr>
<td>Age of Fund</td>
<td>Numeric</td>
<td>Age of fixed since inception</td>
</tr>
<tr>
<td>Standard Creation</td>
<td>Numeric</td>
<td>Fee on redemptions and subscriptions</td>
</tr>
<tr>
<td>Redemption Fee</td>
<td>Numeric</td>
<td>Number of shares traded per redemption and subscription</td>
</tr>
<tr>
<td>Share per Unit Order</td>
<td>Numeric</td>
<td>Market type of fund as indicated on ishares.com</td>
</tr>
<tr>
<td>Market Type</td>
<td>Nominal</td>
<td>Dominant country of the security holdings</td>
</tr>
<tr>
<td>Country Exposure</td>
<td>Nominal</td>
<td>The difference between fund performance and tracked index</td>
</tr>
<tr>
<td>Tracking Error</td>
<td>Numeric</td>
<td>Management fees associated with running the fund</td>
</tr>
<tr>
<td>Management Fees</td>
<td>Numeric</td>
<td>Number of shares per unit order</td>
</tr>
<tr>
<td>Number of Fund Holdings</td>
<td>Numeric</td>
<td>Total number of securities in the fund</td>
</tr>
<tr>
<td>Index Return</td>
<td>Numeric</td>
<td>The index return</td>
</tr>
<tr>
<td>NAV Total Return</td>
<td>Numeric</td>
<td>NAV total return</td>
</tr>
<tr>
<td>Top Holding Weight</td>
<td>Nominal</td>
<td>Percentage weight of the fund's largest security holding</td>
</tr>
<tr>
<td>Beta vs. S&amp;P 500</td>
<td>Numeric</td>
<td>Beta of the fund using the S&amp;P 500 as the base</td>
</tr>
<tr>
<td>12 Month Yield</td>
<td>Numeric</td>
<td>The fund's 12 month yield from 2012 to 2013</td>
</tr>
</tbody>
</table>

The next variable is the family of index being tracked by the fund. Some indices are better known and popular with investors. Since investors may view a well-known index provider like Russell they may be more likely to invest in an ETF whose underlying index is the Russell 2000. Picking the right underlying index for an ETF is essential because it is the main driving factor of the funds performance. ETF funds are expected to closely track the performance of their underlying index.

The age of the fund is also a possible variable in consideration when investors choose to invest. A proven track record of meeting investor expectations would logically attract more investors. But there is also a need to fill investor demand for new frontier and emerging market exposure through the creation of new funds. Much of the ETF market has been saturated with US domestic and European security holdings giving way to more exotic funds entering into more risky emerging markets. Standard creation and redemptions fees are incurred when an authorized participant wishes to buy or sell their ETF shares. The fee is charged by the product provider and its custodian. The size of the fee influences the frequency and liquidity of shares in the market with larger fees inherently decreasing the frequency of trading all else held equal.

The number of shares per unit order is not an attribute of funds fully visible to the investing public. It indicates the number of shares that are created or redeemed by the authorized participant during subscriptions and redemptions in and out of the fund. Authorized participants are large financial institutions such as J.P. Morgan and Goldman Sachs that hold ETF shares to be sold to the investing public and other financial institutions. They have influence over the net flows of funds through the number of shares they wish to hold in their inventory for trading.

Market type is a key variable in marketing ETFs. The market type variable was taken from the iShares website. It indicates the type of market the ETF covers. It is used to give investors an idea of the types of securities in the fund. Investors use it to help balance their investment portfolios. For instance if an investor has a portfolio heavy in the healthcare sector they may be looking to balance their portfolio with REITs or real estate sector funds to gain a more broad exposure. The key to marketing ETFs to investors is by highlighting to investors what the most popular trend is and financial institutions can do this by marketing the sector or market type the fund is in.

Country exposure is very similar to market type as it gives the investor information on the types of securities they can buy into. ETF’s have the advantage over traditional mutual funds by giving investors an instant portfolio diversification. By owning one share of an ETF an investor now how market exposure to the securities and markets from which they are derived. The country exposure variable consists of single or multiple country exposure. To simplify the variable and fund with more than one country is indicated as having multiple country exposure. If the fund only has security holdings from one country the name of the country is noted in the variable e.g. Spain. This is part of the third step in the CRISP-DM process. Country exposure is cleaned and formatted in a way that data mining models could consume the data. Listing out all the countries for funds that have multiple countries could not be used with the data modeling methods in this study.

Tracking error is a calculation done in the third stage of the CRISP-DM process. The tracking error is the difference between the funds performance and the underlying index performance. The most common formula used to calculate the tracking error includes determining the standard deviation of the return difference between the benchmark and the NAV of
the fund. Over n periods the formula for tracking error is calculated below with x being the return of the fund and y being the return on the index. However for this study a more simplistic method is used as data collection for the fund and index returned was done at the start of 2012 to be used as a predictor of net flows. In this case formula 1 can be simplified to formula 2 since only one period is used in the calculation.

\[ \sigma^2 = \frac{1}{(n - 1)} \sum (x_i - y_i)^2 \]  

(1)

\[ \text{Return}_p - \text{Return}_i = \text{Tracking Error} \]  

(2)

where \( p \) = portfolio and \( i \) = index.

Management fees or fund expenses have been a well-documented variable in studies done by Barber et al and others [18]. It is thought that funds with larger expense ratios tend to attract fewer investors as compared to comparable funds with lower expense ratios. Expenses also have a direct effect on fund performance by reducing returns. Therefore management fees are a variable to be considered because it is easily obtained by investors and thought of as being a negative contributor to fund performance. Yet, even with the fee war occurring between ETF providers such as BlackRock and Vanguard management fees should not be a considerable concern to ETF investors as long as the tracking error is held to a minimum. Each ETF provider must devise ways to offset any management fees to ensure the fund tracked the underlying index as closely as possible. The total number of holdings in a fund is a variable comprised of the amount of underlying securities in the fund. The number of holdings in a fund should ideally match the number of holdings in the underlying index and in the same proportion in order to track the index return. An investor may be concerned with a small number of holdings in a fund as it is often harder to keep the right proportion while offsetting management fees. Funds with smaller number of holdings have broader swings in returns. It is also a determining factor in portfolio diversification. Fund holding were calculated at the start of 2012 and are expected to change throughout the year during annual index rebalances but the change is insignificant to this study. Index return and NAV total return are used in the tracking error calculation. The returns themselves may also contribute to investor decisions. As previous studies have indicated, past performance influences future investment decisions by investors. In a study done by Sirri et al performance has a positive contribution to fund flows [4]. The beta versus the S&P 500 is often used in finance to explain the price volatility of funds. It indicates the extent to which a fund’s price will move with respect to the S&P 500 index. Investors use the beta in their risk analysis. A beta higher than 1 indicates the fund prices fluctuates more than the S&P 500. A beta of less than 1 indicates the fund’s price will fluctuate less. More fluctuation means more risk as the price becomes more unpredictable. The 12 Month Yield is the sum of previous 12 month’s dividend and interest income payments over the nearest months Net Asset Value per share as indicated in formula three. It illustrates to the investor the amount of money they can expect to get paid out per dollar value of the fund.

\[ \text{Yield} = \frac{\text{Income}}{\text{NAV}} \]  

(3)

where \( \text{income} \) = the sum of the trailing 12-months interest income and dividend payments.

The sensitivity analysis in this study will reveal the extent these variables influence net flows of ETFs. There are several unused variables in this study. The unused variables include the size of the fund and the degree of competition. The size of the fund is omitted as a possible predictor because it is a direct result of net flows and would skew the results. The degree of competition is also thrown out of this study because of the inability to either assign reasonable numeric value or ordinal value.

### 3.2 Data Mining Methods Employed

There are three data mining methods employed in this study. Each one used the 10 fold Cross Validation to ensure accurate predictive results. The three models used are Alternating Decision Tree Learning [20], Logistic Regression [21]-[22], and Back-propagation Multilayer Perceptron. These three models are used in this study because of their ability to consume numerical and nominal values in predicting the outcome of fund net flows. Each model interprets the data set in a unique way to predict the outcome of the testing data set. The final MLP structure used in this study is illustrated in Figure 2.
4. Results and Discussion

The next step in the CRISP-DM process framework is to review the results of the models used in this study. The results of the data mining models appear in table 2. A comparison of the three classification models is also discussed comparing and contrasting the benefits and disadvantages of the models. The results of the three data mining models show how each models methods of predicting the net share flows produce different outcomes. The results in table 2 are a reconstruction of the results from the WEKA software used in the study. In each model the total set of variables are used as inputs.

The logistics regression model revealed an accuracy of 59.46% for correctly classified instances. 132 out of 222 total instances are correctly classified. The confusion matrix confirms these findings. The confusion matrix is used to visualize the accuracy of predicting models. It illustrates the instances of actual or true positive and negative outcomes as well as the false positive and negative instances. Table 2 shows the confusion matrix on the right for the three data mining methods. The true positives and true negative instances are in green and the false positives and negatives are in red. For the logistic regression model it shows the false positives and false negative instance to be 38 and 52 respectively.

Alternating decision tree learning algorithm results are quite different than the logistic regression model. The overall accuracy of the model is 66.22% using all the input variables. The confusion matrix shows the breakdown of the true and false positive and negative instances from the alternating decision tree learning model. The accuracy indicates the ADT model has a superior algorithm for predicting the outcome of the business problem analyzed in the study.

The artificial neural network used for modeling is the back-propagation multilayer perceptron. In this study the accuracy results of MLP are the least favorable of the three. Only 57.66% of the instances are classified correctly out of 222. The nature of the model makes it difficult to determine where in the decision making process the model failed. Artificial neural networks are said to be like a “black box” with no way of knowing why the final outcome is chosen [23].

The logistic regression model and the back-propagation multilayer perceptron model both are unable to classify the negative instances over 50% indicating no better chance in predicting the negative outcome than of flipping a coin. 52 instances are falsely identified as negative and only 47 are correctly identified as negative. In this category of the confusion matrix the alternative decision tree algorithm is more accurate along with the all other categories. This clearly demonstrates the ADT algorithm is a superior predictor of the outcome. The back-propagation multilayer perceptron model is a better accuracy of true positives versus the logistic regression model yet the false negatives are higher creating an overall lower accuracy for total outcomes.

4.1. Sensitivity Analysis

Sensitivity Analysis results augment the data mining accuracy results from WEKA. Sensitivity analysis shows how individual variables used in the study effect the three models ability to predict an
accurate outcome. Figure 3 depicts the aggregated sensitivity scores of the variables after being normalized across the baseline accuracy of the three models. The sensitivity scores are shown in descending order from left to right in figure 3. Figure 3 shows the variable management fees is the most sensitive and as previously discussed increase the accuracy of the three models when removed. Standard creation redemption fee is the second most sensitive variable across the three models. In two out of the three models, it drastically changed accuracy when removed. Six of the variables show relative similar influences on model results. These six variables include NAV total return, fund exposure, market type, share per unit order, tracking error, and top holding weight percentage.

### 4.2. Managerial Implications

There are several pertinent managerial implications from the results of this study. The first is the model’s ability to accurately predict the net fund flows. Each of the three models could predict outcomes at a rate over 50%. The outcome or output of the models is an answer to a yes/no question. A yes/no question can be answered correctly with no other information accurately 50% of the time and is a basic truth in statistical analysis. Therefore the use of the selected variables to aid in the decision making process of the data mining models indicates there is value to their predicting ability. Financial institutions in business to develop and market exchange traded funds should implement the use of data mining techniques in their decision on fund characteristics. This will allow them to create a financial product that will generate positive inflows. These inflows make the fund a worthwhile investment that financial institutions can rely to offset their costs in the product. Management should also be concerned with the results details and what the findings implicitly say about investor’s reactions to certain fund characteristics. Sensitivity analysis shows management fees are the most influential on investor’s decisions to invest in one particular exchange traded fund over another. Managers of these funds should take this into consideration when they are choosing expense ratios to charge for the operation of the funds. This explains the expense war between the two leading providers BlackRock and Vanguard. The two financial institutions charge significantly different management fees on their funds. Vanguard is using a low fee approach to marketing their funds while BlackRock attempts to re-brand the majority of their funds by changing their names and lowering management fees to be more competitive. The second most sensitive variable is standard creation redemption fees. The fees charged by financial institutions on the creation and redemption process of shares is not public knowledge but does indicate it has a large influence of the outcome of net fund flows. Like that of management fees it is likely to increase fund flows if based on the assumption fees are seen negatively in the eyes of investors. Even so, sensitivity analysis does not indicate how the variables affect the model results but only that they do so disproportionally.

### 5. Conclusions

The application of data mining methods can be used to better the understanding of how fund characteristics influence the decisions of individual investors as they choose to invest their money in one financial instrument over another. They ability of management in financial institutions to gather this information can be a source of competitive advantage and give rise to new product development techniques. In addition the information gained from data mining approaches used in this study can help management focus on marketing the right fund characteristics for exchange traded funds that best contribute to fund inflows. Taking into consideration the results of the three predicting models in the study, alternative decision tree learning algorithms demonstrated accuracy results greater than 68%. As a result, it is recommended that an alternative decision tree learning algorithm be used in business practice in order to predict the success of current and future exchange traded fund development projects.

This study concludes that additional research is still needed to increase the accuracy of modeling outcomes. While 68% clearly shows there are linkages between a number of variables and fund net flows it does still give much room for improvement. This study is conducted during a 10 week period of
which only 3-4 weeks was spent on data collection. Further research can be done on other possible variables given more time. The marketing budget, broker incentives, and degree of competition are future variables to be considered in future data mining models but could not be included in this study due to their complexity and availability of creditable sources.

The results of this study support the need for larger data set of variables to be tested. Previous research has been done on individual variables with linear relationships. Some of the research is contradicting and lacking unified findings. The use of data mining methods has shown to provide useful comparisons between classification models. The results can be linked to current market practices as in the case of management fee reductions in the exchange traded fund industry. Most importantly this study is one of the first of its kind on examining the predictive power of fund characteristics on the fund net share flows of exchange traded funds. The vast majority of prior research centered on the mutual fund industry as a whole and conducted studies on traditional mutual funds aiming to outperform the market. These studies lacked a strong connection to the exchange traded fund universe with several important and basic differences among them. This study has successfully helped to lesson that gap and provide useful insight into why investors choose one exchange traded fund over another.

6. References


