A Bankruptcy Risk Factor

Jesse Neumann*

Rutgers University

Draft November 2021
Most recent version Here

Abstract

This paper introduces a factor based on an estimated probability of bankruptcy — a measure of the risk a typical investor will lose their investment, or the cost of insuring that investment. Using an underlying model of firm bankruptcy built as a sequence of two random forests, I demonstrate my bankruptcy risk factor has predictive power in equity, bond, option and credit default swap markets earning statistically significant monthly returns of 0.23%, 0.15%, 1.97% and 1.04%, respectively, in all four markets. In markets with existing common factors I find statistically significant alpha with respect to these factors.

1 Introduction

Explanations for why some assets receive higher returns than others is the fundamental question in empirical asset pricing. From an equilibrium theory perspective the answer is clear — higher returns are demanded by investors for holding assets with higher levels of risk. Put differently, there exists a pricing kernel such that risk-adjusted prices in terms of next period payoffs are the same. However, empirical results do not always agree with this well-studied theoretical answer. In the equity space, the illiquidity factor of Amihud and Mendelson (1986), the size factor of Fama and French (1993) and the credit default swap term factor of Friewald et al. (2014) all indicate that higher levels of risk are associated with higher returns. At the same time, the cash flow volatility factor of Huang (2009), the profitability factor of Fama and French (2015) and the return on equity factor of Hou, Xue and Zhang (2015) indicate lower risk is associated with higher returns.

Although factors based on various indirect proxies for risk ultimately disagree on whether

*Graduate student, Department of Economics, Rutgers University. Email: jn495@economics.rutgers.edu.
I am indebted to Bruce Mizrach for his comments, suggestions and overall guidance through this research process. I have also received helpful comments from participants at the Rutgers University econometrics seminar. All errors are my own.

higher returns are compensation to investors for holding higher levels of risk, more direct measures of risk produce the same — perhaps surprising — result. Firms with lower levels of risk are associated with higher average returns. This distress risk literature includes Griffin and Lemmon (2002) and Dichev (1998) who propose portfolios sorted on Ohlson's O-Score (Ohlson, 1980). Dichev (1998), who also proposes portfolios sorted on Altman's Z-Score (Altman 1968). Vassalou and Xing (2004) who propose portfolios based on default risk as implied by Merton's (1974) option pricing model. Campbell, Hilscher and Szilagyi (2008) who proposes portfolios based on a logit model estimated failure probability. Avramov et al. (2009) who propose portfolios based on credit downgrade events, and Asness, Frazzini and Pedersen (2019) who proposes portfolios based on a measure of firm quality. Despite the different definitions of distress, the direction of the sort is always the same — firms with lower levels of distress (either lower bankruptcy or default probability, higher credit ratings or safer) earn higher expected returns.

The increasing demand by investors for safe, or low risk equity assets during periods of market downturn is a well-known phenomenon. However, this flight to safety[1] during market downturns alone is unlikely to be driving the equity factor results since these periods occur infrequently. To reconcile the equilibrium theory and empirical distress risk results Asness, Frazzini and Pedersen (2019) present a dynamic model based on residual income which shows how flight to safety also reasonably occurs during normal market conditions. In this model fundamental firm value scaled by book value increases with safety. A higher fundamental value compared to book value indicates the firm is undervalued and future expected returns should be higher.

While equity portfolios are by far the most frequently studied, the equilibrium theory *asset* pricing result should hold for all *assets*, not just equities. More than $500 billion in bonds are traded each day. Additionally, the large trading volumes in options and other derivatives show these assets are not redundant with respect to their underlying equities — it is thus important to study them in adequate detail (Almeida and Freire, 2021). Fons (1987, bonds sorted by credit rating), Elton et al. (1995, bonds sorted by default risk) and Bai et al. (2019, bonds sorted on credit rating) all demonstrate that higher risk bonds are associated with higher expected returns. This difference in the direction of increasing returns between bonds and equities is driven by the coupon payments of bonds. Lower rated (higher risk) bonds generally require larger coupon payments, and since accrued interest is built into bond prices, bonds with higher coupon rates (i.e. faster accruing interest) have higher rates of return.

Cao et al. (*forthcoming*) demonstrate that option returns are increasing in bankruptcy risk as measured by the Altman Z-Score[2]. This is consistent with the use of options as levered bets. Conversely, Friewald et al. (2014) show that the credit default swap (CDS) term structure is positively correlated with equity returns. This implies we should expect higher risk to be associated with lower returns to holding CDS contracts. My results will

[1] Also termed flight to quality.

[2] They find that a higher Z-Score, which implies lower bankruptcy risk, is associated with lower returns.

confirm this is true.

In this paper I introduce a new factor based on the probability of firm bankruptcy in each of the four markets discussed above — equity, bond, option and CDS. This paper differs from previous studies sorting firms on measures of distress in a number of key dimensions. First, I use the risk of bankruptcy as the sorting characteristic rather than other forms of distress risk. Bankruptcy risk is less subjective (it is a discrete, easily definable and observable event) and the legal process of going through bankruptcy can trigger larger affect on asset values than other forms of distress risk. Second, I develop a new bankruptcy risk model instead of using an existing model — such as Altman's Z-Score or Ohlson's O-Score — like previous studies have done. These models are decades old and it is likely the risks facing firms have evolved over time. I will provide evidence for this throughout the paper.

Third, the bankruptcy risk model utilizes machine learning methods, being built as a sequence of two random forests. The first random forest selects a parsimonious model in a data driven way using the feature importance metric of Breiman (2001), and the second random forest trains the prediction model. The bankruptcy prediction literature has proposed hundreds of potential predictors and utilizing a random forest in this way allows me to avoid having to take an *ex ante* stance on which predictors are the most important. Parsimony turns out to be important not only when interpreting the contribution of individual inputs, but also because a model using the full set of predictors overfits when applied out-of-sample. Section 2 discusses the bankruptcy risk model is more detail. While applying the latest machine learning techniques to bankruptcy prediction is not new, using the output of a machine learning based bankruptcy prediction model to form characteristic sorted portfolios has not yet been attempted in the literature. I will demonstrate that the increased accuracy achieved by the random forest compared to Altman's Z-Score[3] and a weighted least squares model using the same five variables as the Z-Score leads to more profitable portfolios.

Lastly, I use my bankruptcy risk model to form characteristic sorted portfolios in four markets. This differs from previous studies which generally introduce factors in only a single market at a time (see e.g. Harvey and Liu (2019) for a census of equity factors, Bai et al. (2019) for a default risk bond factor and Bakshi and Kapadia (2003) for an implied volatility option factor). While there are papers highlighting anomaly factors which have explanatory power in multiple markets (see e.g. Asness et al. (2013) which includes bonds, currencies and commodity futures in addition to stocks; Chordia et al. (2017) which includes stocks and bonds; Frazzini and Pedersen (2014) which includes bonds and futures in addition to stocks; Moskowitz et al. (2012) which includes curencies, commodities and bonds in addition to stocks; and Brooks et al. (2018) which includes options in addition to stocks), this paper is the first to simultaneously introduce a factor with the same construction in equity, bond, option and CDS markets. To my knowledge this is the first paper to provide initial evidence

[3]I choose the Z-Score model as a benchmark because, although it was proposed in 1968, it is still one of the most widely used models by academics and practitioners. Edward Altman recently partnered with the Kroll Bond Rating Agency applying his expertise, including use of the Z-score, to the analysis of corporate default risk. Additionally, the original Z-score paper has 20,553 citations on Google Scholar as of October, 2021.

of a factor structure in the CDS market.

The results indicate that my estimated probability of bankruptcy holds predictive power for equity, bond, option and CDS markets. My long-short factor — which I title SSD_i (Safe Subtracting Distressed, for $i \in (e, b, o, c)$) for the equity, bond, option and CDS factors — produces statistically significant monthly returns of 0.23%, 0.15%, 1.97% and 1.04% in each respective market. Returns to SSD_e are also higher during crisis periods, consistent with the flight to safety phenomenon.

Notably, the variables selected by the initial random forest for use as bankruptcy predictors — return on assets, scaled and unscaled net income, scaled pretax income and current assets scaled by current liabilities — are largely absent from the factor literatures in all of the markets studied here. Interestingly, all five of these variables are accounting ratios or income statement items even though there are market based variables available for selection. In their census of the factor zoo, Harvey and Liu (2019)[4] find only two factors out of over 500 are generated by sorting on any of these five variables. Likewise, only three of the 319 factors included in Chen and Zimmerman (*forthcoming*) and three of the 94 factors included in Gu, Kelly and Xiu (2020) are generated by sorting on any of these five variables. Furthermore, only one factor — generated by sorting on Piotroski's (2000) distress measure which includes return on assets — is deemed important by the machine learning models of Gu, Kelly and Xiu (2020). This paper therefore also demonstrates the ability of standard financial ratios to predict asset returns when combined in the correct way.

The equity, bond and option factors also produces statistically significant alpha when regressed on popular existing factors. For SSD_e this includes the CAPM and the Fama-French five-factor model. SSD_e also contributes to the explanation of the cross-section of returns when included in the factor zoo using the two pass lasso procedure of Feng, Giglio and Xiu (2020). For SSD_b this includes regressions on the excess bond market, six common bond factors (excess bond market, two measures of illiquidity, term structure, default risk and momentum) and the Fama-French equity factors. For SSD_o this includes regressions on a factor formed on Altman's Z-Score, six common option factors (implied volatility, illiquidity, size, the difference between implied and realized volatility, idiosyncratic volatility and book-to-market) and the Fama-French equity factors. To my knowledge there are no other CDS factors for me to compared SSD_c to, but my hope is SSD_c could be used in future studies as an explanatory variable in spanning regressions.

The remainder of this paper is organized as follows. Section 2 details the creation of the bankruptcy risk measure. Section 3 presents the data sources used to create the bankruptcy predictor variables and the equity, bond, option and CDS factors. Section 4 presents the resulting performance of the equity, bond, option and CDS factors and finally section 5 concludes.

[4]The census can be found here: `https://docs.google.com/spreadsheets/d/1mws1bU56ZAc8aK7Dvz696LknM0Vp4Rojc3n61q2-keY/edit?usp=sharing`

2 Bankruptcy Risk Measure

This section details the two-step bankruptcy risk model used to create SSD_i (for $i = e, b, o, c$). There are two purposes in creating a new model of firm bankruptcy. The first is to find a parsimonious set of predictors without having to take a stance, ex ante, about which among the extant predictors are most important. The second is that — insofar as the accuracy of the bankruptcy risk measure matters for the quality of SSD_i (for $i = e, b, o, c$) — it is important to consider the most relevant risk factors facing firms today. Previous bankruptcy risk models were tuned to the risks facing firms when those papers were published and may not be reflective of the risks facing the contemporary firm. One of the benefits of this measure is that — unlike the Z-score and related measures — it is directly interpretable as the probability a firm will declare bankruptcy within 12 months.

Before detailing how the bankruptcy risk measure was developed, it is important to acknowledge that the legal process of filing for bankruptcy is not the only possible measure of firm distress. Failure to pay exchange listing fees, loan default, raising capital for the express purpose of continuing operations (Jones and Hensher, 2004) and reductions in dividends (DeAngelo and DeAngelo, 1990), among others, have also been used as measures of firm distress. However, many of these measures have some degree of subjectivity or could be the result of errors or strategic decisions unrelated to distress. For this reason I choose to measure firm distress as a firm initiating the bankruptcy process.

Perhaps the biggest difficulty for any model of corporate bankruptcy is that the formal filing of bankruptcy by a publicly traded firm is rare. Even for studies using a small number of pre-selected bankruptcy predictors, the number of bankrupt firms in the sample can be very small. Altman (1968) uses a sample of only 33 bankrupt firms. Ohlson (1980) uses a sample of 105 bankrupt firms. Lennox (1999) uses a sample of 90 bankrupt firms. This problem is exacerbated here because to be included in the dataset each firm must have non-missing entries for all variables necessary to create the full set of predictors instead of the much smaller number of variables used in previous models. This reflects my competing goals of choosing — in a data driven way — from among the largest possible number of bankruptcy predictors introduced in the literature and including the maximum number of bankruptcies possible in order to obtain the most robust model.

There are 457 bankruptcies of publicly traded firms listed in the CRSP/Compustat merged database between 1990 and 2019. After accounting for missing data, requiring multiple observations for each firm (necessary because some predictors require differences between adjacent observations or moving averages), and lagging the dependent variable I was left with 71 bankrupt firms. Table 1 details this process. Since the year bankruptcy is initiated occurs only once, there are only 71 bankrupt firm-year observations. However, there are 54,389 firm-year observations for non-bankrupt firms making this a highly imbalanced classification problem. The 0.13% bankruptcy rate in my sample is of the same magnitude as the 0.76% bankruptcy rate found in Zmijewski (1984) and is reasonably less considering the

overall decrease in the corporate bankruptcy rate since the 1980s[5]. The similarity between the degree of class imbalance provides a level of validity to the composition of my dataset. The econometric and machine learning literatures both acknowledge the problems associated with highly imbalanced classes (see e.g. Fernàndez et al., (2018) and King and Zeng (2001)) although there is no consensus on the best way to resolve the problem.

Table 1: Bankruptcy Predictor Data Sources

Database	Variables	Firm-Year Obs	Bankruptcies
Entire Compustat Universe	**1990 - 2020**	**388,376**	**457**
Compustat Fundamentals	Accounting	83,845	171
Compustat Names	Industry	83,845	171
Compustat Segments	No. Business Segments	83,845	171
CRSP	Market	71,270	107
WRDS Ratio Suite	Financial Ratios	64,645	96
One Year Lag		54,460	71
K-Nearest Neighbor Match		**142**	**71**

Note: This table summarizes the data sources used to create the bankruptcy predictor variables. Beginning at the top of the table, the decreases in Firm-Year Obs and Bankruptcies indicates the number of observations lost due to missing data as each subsequent database was added to the dataset.

Consistent with the use of machine learning for modeling bankruptcy, I also use machine learning to address the class imbalance, utilizing K-nearest neighbors to match a bankrupt firm with a non-bankrupt firm based on size (i.e. market capitalization). Size matching in various forms is common in finance and financial economic research. Bennett and Wei (2006) match on firm size when comparing stocks listed on the NYSE and NASDAQ. Many studies (see e.g. Novy-Marx (2013)) scale variables by either total assets or market capitalization, the purpose being to remove the influence of firm size on comparisons. Altman (1968) also uses a type of size matching in his original Z-Score analysis where the non-bankrupt firms were chosen as a stratified random sample from a group of firms with the same range of market capitalization as the bankrupt firms.

Matching in this way does present a potential problem. Small firms file for bankruptcy at a much higher rate than large firms so matching on firm size means the sample used to train the bankruptcy risk model has a smaller average market capitalization than the unconditional average among publicly traded firms. This highlights the importance of choosing a parsimonious model to avoid overfitting this lower market capitalization training sample.

Using the K-nearest neighbors matched sample of 142 firm-years (71 bankrupt and 71 non-bankrupt) the two-step random forest procedure is implemented as follows.

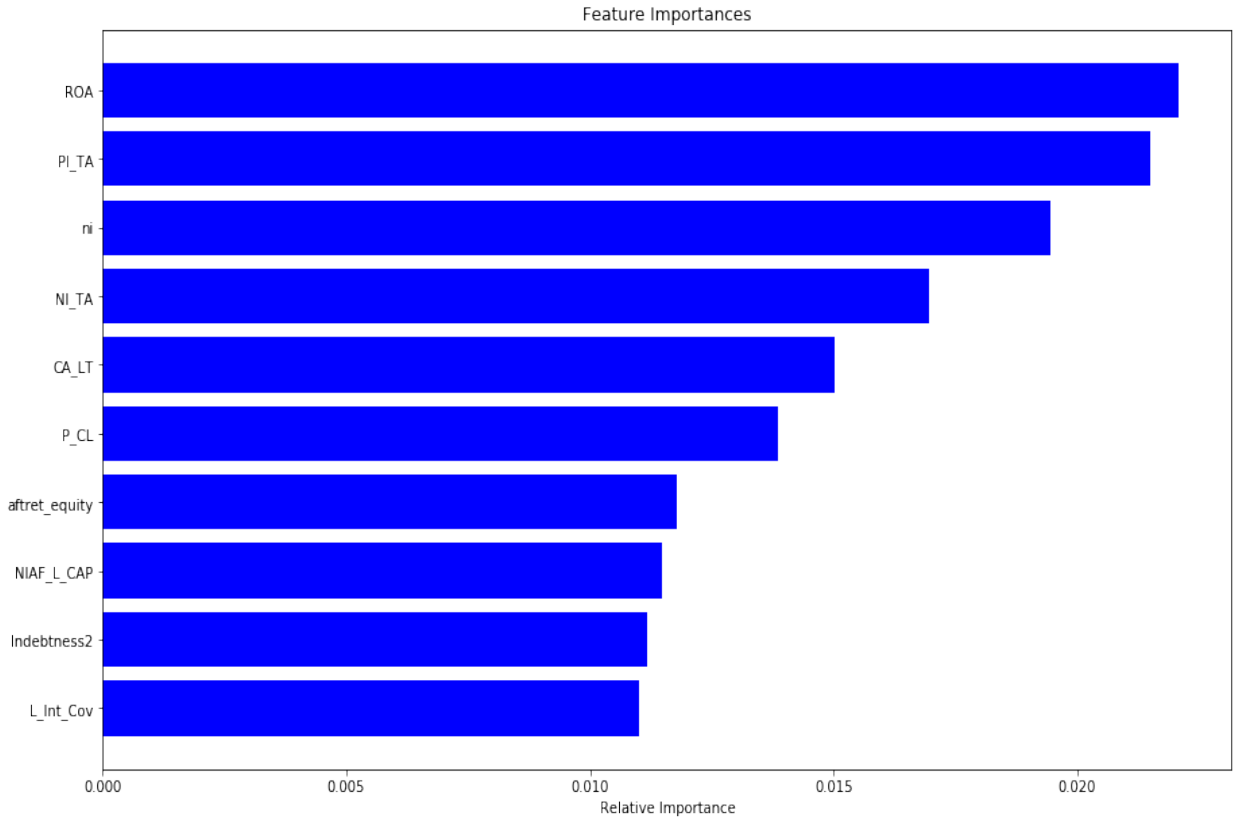
1. Using the full list of bankruptcy predictors, train a random forest and evaluate each feature's importance using the method of Breiman (2001).

[5]Source: Administrative Office of the U.S. Courts - accessed here `https://tradingeconomics.com/united-states/bankruptcies`

2. Select the desired number of features from the feature importance ranking. In this paper I select features until there is no longer a large decrease in importance between subsequent features. Train a second random forest using only those selected features.

I utilize random forests for two reasons. First, in the setting of hundreds of independent variables and limited observations, the random forest allows for the measurement of non-linear relationships while simultaneously being robust to outliers and guarding against overfitting. Second, since random forests are a nonparametric ensemble learning method it can be reasonably trained using the 142 observations I have, compared to a neural network which requires a much larger dataset to obtain robust results.

Figure 1: Random Forest Feature Importance



Note: this figure presents the relative feature importance of each bankruptcy predictor based on the importance metric of Breiman (2001). *ROA* is return on assets, *PI_TA* is pretax income scaled by total assets, *ni* is net income, *NI_TA* is net income scaled by total assets, *CA_LT* is current assets scaled by total liabilities, *P_CL* is gross profitability scaled by current liabilities, *aftret_equity* is the after-tax return on shareholder equity, *NIAF_L_CAP* is net income after taxes scaled by the lag of capital investment, *Indebtness2* is the square of liabilities scaled by liabilities and shareholder equity, *L_Int_Cov* is the log of the interest coverage ratio.

An additional benefit of using this two-step procedure is that it provides a framework by which the bankruptcy risk model can be updated to both include newly proposed predictors and ensure the most relevant predictors are included in the model at any point in time.

For this study the first forest is grown using 100 trees able to use 16 randomly selected variables at each node to partition the data following the $m_{try} = \sqrt{p}$ rule (Probst et al., 2019). That is, the number of randomly selected variables for use in the optimization procedure at each node is equal to the square root of the total number of variables in the data set. Entropy is used as the criterion function at each node and trees are allowed to grow until there is only observation in each terminal node[6]. Figure 1 displays the ten most important features according to the feature importance measure of Breiman (2001).

If Figure 1 was extended to include all 241 bankruptcy predictors the same pattern would continue — the difference between the remaining adjacent predictors is very small. Although I choose the first five predictors as inputs for the second random forest, it is evident from Figure 1 that there is also a large relative decrease in importance after gross profitability scaled by current liabilities (P_CL). This ratio was omitted from the final model because it did not increase model accuracy in the test set and parsimony is important to avoid overfitting. Parsimony also makes the impact of each individual predictor on the probability of bankruptcy easier to interpret.

The selection of these five ratios are noteworthy for two reasons. First, although market based predictors of bankruptcy have existed since Beaver (1968), there has been a recent shift from once popular accounting ratios to more market based variables as predictors of bankruptcy due to their enhanced perceived accuracy. Hillegeist et al. (2004) find market based bankruptcy predictors produce more accurate bankruptcy risk models, while Agarwal and Taffler (2008) find little difference in the predictive accuracy of market based and accounting ratio based bankruptcy risk models. My findings add support for accounting based bankruptcy risk models, even in the presence of market based predictors.

The second reason is because portfolios sorted on any of these five characteristics are largely absent from the discussion of the factor zoo and are rarely, if ever, selected as important factors by machine learning methods. In their census of the factor zoo, Harvey and Liu (2019) document only two out of 524 characteristic sorted portfolios use any of the five variables selected by the random forest — both of which were return on assets. This compares to 34 factors created on the market return, or some transformation thereof. In their influential paper on the use of machine learning in asset pricing, Gu, Kelly and Xiu (2020) include only three portfolio sorts based on the variables selected by my random forest out of 94 total characteristics. Additionally, their eight machine learning methods only select one — Piotroski's (2000) f-score which includes return on assets in its construction — when ranking the 20 most important predictors. In fact, their machine learning methods indicate the 10 most important predictors are all market variables (price trends, liquidity and volatility).

To train the second random forest — which produces the estimated probability of bankruptcy

[6]Although they are not required to grow until there is only one observation in each terminal node. If there is more than one observation at a node, all belonging to the same class, then there is no need to further partition the data at that node.

— I split the data into a training set of 122 observations (61 bankrupt firms and 61 non-bankrupt firms) and a test set of 20 observations (10 bankrupt firms and 10 non-bankrupt firms) to evaluate the model's performance. This forest is grown using 1,000 trees able to choose from only two of the five selected predictors from the first random forest at each node. To evaluate the performance of the random forest against previous bankruptcy prediction models I compare the predictive accuracy of the random forest on the test set with the the original Altman Z-Score as well as a weighted least squares re-estimation of the coefficients used in Altman's Z-Score.

If the risk factors facing firms do, in fact, evolve over time as hypothesized we would expect to see two things. First, the coefficients on the inputs used by Altman (1968) should be different from the coefficients on a model with the same inputs, but estimated using my updated sample of bankrupt and non-bankrupt firms. Second, the random forest should select different predictors and should be more accurate than both Altman's Z-Score and the re-weighted Z-Score. Figure 1 has already shown the random forest selects alternative predictors than those used in Altman (1968). Equations 1 and 2 below compare the coefficients of Altman's Z-Score and the re-estimated Z-Score, while Table 3 compares model accuracy.

$$Z = 0.012X_1 + 0.014X_2 + 0.033X_3 + 0.006X_4 + 0.999X_5 \tag{1}$$

$$Z_{WLS} = -0.159X_1 - 0.021X_2 + 0.861X_3 + 0.001X_4 - 0.179X_5 \tag{2}$$

where X_1 is working capital scaled by total assets, X_2 is retained earnings scaled by total assets, X_3 is earnings before interest and taxes scaled by total assets, X_4 is market value of equity scaled by total liabilities and X_5 is sales scaled by total assets. The change in sign and magnitude of the coefficients from equation 1 and equation 2 implies an omitted variable bias since it is unlikely a higher rate of sales per asset (X_5, for example) would mean a higher risk of bankruptcy. This indicates there are missing risk factors from the regression — that is — risk factors are different today than they were in 1968. Further evidence for this assertion is presented in Tables 2 and 3.

Table 2 presents the mean values in my updated sample of both the original five ratios used by Altman (1968) and the five ratios selected by the random forest, along with the t-statistics and p-values of difference in means tests between the bankrupt and non-bankrupt groups. Not only are the difference in means tests all statistically significant for the random forest selected variables, but most of the original Altman ratios are statistically indistinguishable from each other. Additionally, two of the Altman variables — earnings before interest and taxes scaled by total assets and sales scaled by total assets — have group means which are counter to what we would expect from bankrupt and non-bankrupt firms. It is unlikely in the full universe of publicly traded firms that bankrupt firms have higher sales per asset than non-bankrupt firms, and higher retained earnings per asset than non-bankrupt firms. The fact this is observed in the sample used to update the coefficients of Altman's Z-score means even a bankruptcy risk model with these updated coefficients is unlikely to predict bankruptcy well out-of-sample.

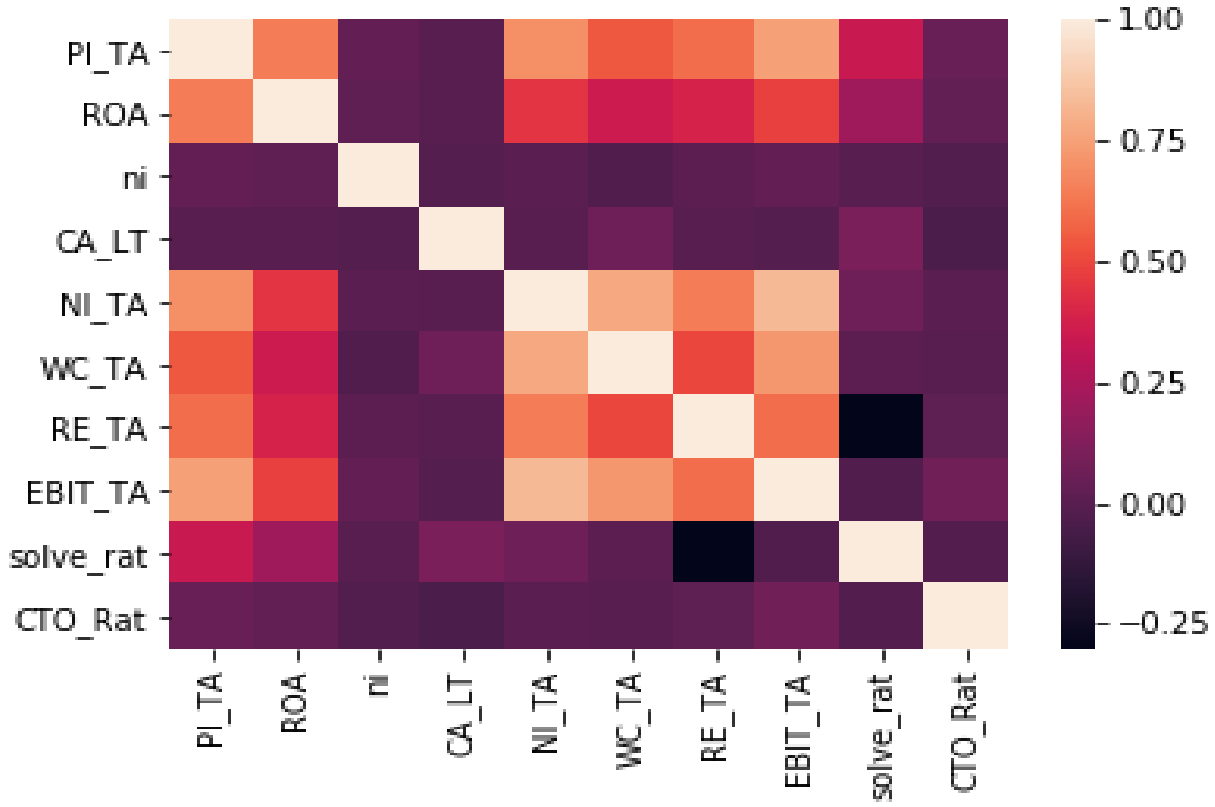
Table 2: Comparison of Bankrupt and Non-Bankrupt Ratios

	Bankrupt	Non-Bankrupt	t-stat	p-value
Altman Z Variables				
$\frac{\text{Working Capital}}{\text{Total Assets}}$	0.142	0.352	3.61	0.0004
$\frac{\text{Retained Earnings}}{\text{Total Assets}}$	-0.604	-0.919	0.51	0.6106
$\frac{\text{Earnings Before Interest and Taxes}}{\text{Total Assets}}$	-0.191	-0.043	-2.03	0.0446
$\frac{\text{Market Value of Equity}}{\text{Total Liabilities}}$	6127	9886	0.59	0.5518
$\frac{\text{Total Sales}}{\text{Total Assets}}$	1.429	1.382	-0.24	0.8043
Random Forest Selected Variables				
$\frac{\text{Income Before Extraordinary Items}}{\text{Total Assets}}$	-0.29	-0.069	2.85	0.0049
$\frac{\text{Pretax Income}}{\text{Total Assets}}$	-0.301	-0.689	2.86	0.0047
Net Income	-28.2	-1.76	2.28	0.0241
$\frac{\text{Net Income}}{\text{Total Assets}}$	-0.299	-0.089	2.74	0.0068
$\frac{\text{Current Assets}}{\text{Total Liabilities}}$	1.38	2.73	2.23	0.0271

I demonstrate in Table 3 this is indeed true. However, the correct sign for the difference in means tests is only half the battle. For the random forest to predict well, it must be the case that the random forest selected variables are independent of — that is, uncorrelated with — these problem Altman variables. Figure 2 confirms this independence. Net income and current assets scaled by current liabilities are uncorrelated with all eight remaining predictors. Likewise, all five of the random forest selected predictors are uncorrelated with the capital turnover ratio — one of the problem Altman ratios. Overall the correlation matrix indicates my bankruptcy prediction model has the potential to be more accurate than both the Altman Z-score, and a regression model using the Altman ratios as explanatory variables.

Table 3 confirms the random forest model is the most accurate of the three models compared. Panel A presents in-sample results and confirms the unique ability of machine learning methods to fit models in-sample. The bootstrapped standard error of the random forest in-sample is 0, indicating the model is able to perfectly fit the training data in all 1,000 bootstrapped samples. The results in Panel B — which displays each model's predictive accuracy using the 20 observation test set — are also indicative of evolving risk factors. When applied to a broad range of modern firms the Z-score predicts bankruptcy only marginally better than a naive guess. When only the coefficients are updated using weighted least squares (weighting by market capitalization) the model's predictive accuracy increases consistent with increased relevance for contemporary risk factors. Further improvements in predictive accuracy occur when new predictors are selected by the random forest, again consistent with the notion of evolving firm risk.

Figure 2: Correlation Between Altman and Neumann Bankruptcy Predictors



Note: This figure presents a heatmap correlation metrix for the ten variables used in my random forest bankruptcy risk model and Altman's Z-Score. *ROA* is return on assets, *PI_TA* is pretax income scaled by total assets, *ni* is net income, *NI_TA* is net income scaled by total assets, *CA_LT* is current assets scaled by total liabilities, *WC_TA* is working capital (current assets minus current liabilities) scaled by total assets, *RE_TA* is retained earnings scaled by total assets, *EBIT_TA* is earnings before interest and taxes scaled by total assets, *solve_rat* is market equity scaled by total liabilities, *CTO_Rat* is the capital turnover ratio (sales scaled by assets).

In this, admittedly small, test sample the random forest bankruptcy model correctly predicts 80% of both bankruptcy and non-bankrupt firms. This is an impressive feat considering market capitalization — a well-known predictor of bankruptcy risk — is not available to help discriminate between high and low bankruptcy risk firms. Bootstrapped 95% confidence intervals are in parentheses.

Because of the number of variables necessary to construct the full set of extant predictors and the requirement that there be no missing data, there were not many observations available to test these three models out-of-sample. As a further robustness test for the ability of each model to predict corporate bankruptcy, Panel C of Table 3 displays each model's predictive accuracy using an expanded test set of 313 firms — 144 bankrupt firms and 169 non-bankrupt firms. This expanded test set was constructed in the same way as the original 144 observation dataset, but instead of deleting missing observations for all of the variables necessary to construct the full set of predictors used in the first random forest, only missing

observations for the variables selected by the first random forest are deleted.

Table 3: Bankruptcy Model Predictive Accuracy

Panel (A)			
Model	**Accuracy**	**FPR**	**FNR**
Original Altman	53.3%	6.4%	86.9%
	(52.0%, 54.6%)	(5.3%, 7.6%)	(85.8%, 88.0%)
Weighted Least Squares	59.8%	47.5%	32.8%
	(59.4%, 60.2%)	(46.8%, 48.2%)	(32.0%, 33.6%)
Random Forest	100%	0%	0%
	(100%, 100%)	(0%, 0%)	(0%, 0%)
Panel (B)			
Model	**Accuracy**	**FPR**	**FNR**
Original Altman	55.0%	40.0%	50.0%
	(52.4%, 57.6%)	(35.8%, 44.2%)	(46.4%, 53.6%)
Weighted Least Squares	60.0%	50.0%	30.0%
	(59.5%, 60.5%)	(49.2%, 50.8%)	(29.2%, 30.8%)
Random Forest	80%	20%	20%
	(79.2%, 80.8%)	(18.9%, 21.1%)	(18.9%, 21.1%)
Panel (C)			
Model	**Accuracy**	**FPR**	**FNR**
Original Altman	58.8%	40.2%	42.3%
	(58.6%, 59.0%)	(40.0%, 40.4%)	(42.0%, 42.6%)
Weighted Least Squares	55.6%	24.8%	67.3%
	(55.4%, 55.8%)	(24.3%, 25.3%)	(66.8%, 67.8%)
Random Forest	68.7%	32.5%	29.9%
	(68.6%, 68.8%)	(32.2%, 32.8%)	(29.6%, 30.2%)

Note: This table presents statistics related to the accuracy of the bankruptcy prediction models. Panel (A) presents the in-sample accuracy, Panel (B) uses a test set of 20 observations, Panel (C) uses a test set of 313 observations. Accuracy is the total percentage of test set observations classified correctly. FPR is the false positive rate — the percentage of non-bankrupt firms classified as bankrupt. FNR is the false negative rate — the percentage of bankrupt firms classified as non-bankrupt. The bootstrapped 95% confidence interval is in parentheses.

It is to be expected that the predictive accuracy of the random forest decreases somewhat in this extended sample. The firms which comprised the original dataset were the largest of the bankrupt firms since smaller firms are more likely to have missing data leading to their removal. As observations were re-added to form the test set used in Panel B, a higher degree of heterogeneity entered the test set as smaller firms were added back. Despite this added heterogeneity the random forest prediction model performs well, correctly classifying nearly 70% of firms, well above Altman's Z-Score and the WLS models. While a nearly 30% false negative rate may be troubling, the random forest still outperforms other industry standard models. Additionally, the conservative manner in which the model was constructed — using only five accounting ratios — ensures the random forest does not overfit when applied to the

entire universe of publicly traded firms. I will demonstrate in section 4 that this conservative model specification still leads to trading strategies with increased returns relative to existing factors.

The power of the random forest bankruptcy probability comes from the fact that despite the relatively small sample size used to train the model, it can be used to make predictions for almost all publicly traded companies. This is because the five inputs selected all use common balance sheet and income statement items which are regularly reported by firms with listed stocks, bonds and derivatives. Contrast this with a predictor such as the R&D intensity measure of Franzen et al. (2007) which requires a research and development variable available for only 32.6% of firm-year observations in Compustat and the benefits become clear.

To this point the random forest has been evaluated on a binary outcome — it either predicts a firm will go bankrupt or it predicts a firm with remain solvent. Random forests make this binary assignment using majority rule. If more than 50% of the trees in the forest predict the firm will go bankrupt than the random forest classifies the firm as bankrupt. However, this binary classification is not very useful when sorting a continuum of firms on bankruptcy risk. I convert this binary classification into a probability by recording the percentage of trees in the forest that predict bankruptcy. For example, if 800 of 1,000 trees predict a firm will go bankrupt in the next 12 months I say that firm has an 80% probability of bankruptcy. It is this probability of bankruptcy I use to sort firms in the remainder of this paper.

3 Data

This section documents the data used and the locations where the data were obtained for both the bankruptcy prediction model and equity, bond, option and CDS prices and characteristics. All data were obtained from Wharton Research Data Services (WRDS) and the Bloomberg Terminal.

The full set of bankruptcy predictors requires data from Compustat Fundamentals (table COMP.FUNDA) for accounting characteristics, Compustat Names (table COMP.NAMES) for industry codes, Compustat Segments (table COMP.SEG_TYPE) for primary business segment, CRSP (table CRSP.MSF) for price and issuance data, and WRDS ratio suite (table WRDSAPPS.FINRATIOFIRM). Compustat, CRSP and the WRDS financial ratio suite were linked using the CRSP-Compustat linking table (CRSP.CCMXPF_LINKTABLE). Of the 241 total predictors constructed, 221 were proposed by the extant literature while I augment these with 20 ratios from the WRDS financial ratio suite. Table 1 in the previous section details the number of observations lost as each database was added. The formation of all predictors proposed by the literature is detailed in Appendix C.

SSD_e is generated using monthly returns data from CRSP from July 1962 to December 2019. Only stocks with share codes 10 or 11 and exchange codes 1, 2 or 3 are included in the equity factor formation. That is, only common stock of U.S. based companies which trade on the NYSE, NASDAQ or American stock exchanges are included. Following Gu,

Kelly and Xiu (2020) I do not impose any price or industry filters. They argue this practice became common in large part because the asset pricing literature found it difficult to model the return behavior of these firms. In unreported results I impose various forms of these filters (i.e. $1 price filter, $5 price filter, etc.) and find similar results.

SSD_b is generated using data from the Financial Industry Regulatory Authority's (FINRA) Enhanced Trade Reporting and Compliance Engine (TRACE). The enhanced TRACE database contains intraday trade by trade data for corporate bonds in the United States. The enhanced TRACE differs from the standard TRACE only by not truncating the volume of trades at $1 million for high yield bonds and $5 million for investment grade bonds. For the period July 2002 to September 2020 bond returns (inclusive of accrued interest) are obtained, already calculated, from WRDS. These returns are generated after applying the filtering procedure of Asquith, Covert and Pathak (2019)[7] and Dick-Nielsen (2009, 2014). Specifically, the following trades are removed: cancelled and updated trades, trades involving bonds with variable rate coupons, trades of bonds issued by firms covered by rule 144a, and trades of bonds other than corporate bonds[8]. Returns are winsorized at the 1% level to mitigate the effect of any data errors. Table 4 summarizes the bond data cleaning process.

Table 4: Bond Data Filtering Procedure

Filter	Observations	Deleted	% Deleted
Total Trades (Enhanced TRACE, 2002 - 2020)	**272,206,673**		
Only last 5 trading days	51,341,154	220,865,519	81.14%
Cancelled and withdrawn trades	49,711,282	1,629,872	0.60%
Aggregate to daily and keep last trading day	5,312,388		
Total Firm-Month Observations	**5,312,388**		
Variable rate coupon	3,741,408	1,570,980	29.57%
Rule 144a	3,734,105	7,303	0.14%
Corporate bonds only	2,334,856	1,399,249	26.34%
Time to maturity greater than 1 year	2,148,945	185,911	3.50%
Final firm-month observations	**2,148,945**		

Default returns (the bond equivalent of delisting returns in CRSP) are generated following Cici, Gibson and Moussawi (2017). Specifically, investment grade bonds are given a return of -17.67% the month of default and high yield bonds are given a return of -40.17% the month of default. That default returns are not -100% is reflective of the fact that defaulted bonds are still tradeable and holders of defaulted bonds often received a strictly positive percent of the principal back. I extend this return series to December 2020 using enhanced TRACE trade data and the same filtering procedure used by WRDS.

SSD_o is generated using data from OptionMetrics from January 1996 to December 2020. Only call options are used since they are less likely to be affected by early exercise (Christoffersen et al. (2018)) and I want the benefits of my options factor to be driven by changes in

[7]Revised in 2019, originally submitted in 2013.

[8]Double-counting dealer trades used to be a problem when Dick-Nielsen (2014) first published his paper, but FINRA has since removed this double-counting from the data.

option prices, not exercise of the option. Additionally, shorting call options serves a similar purpose as put options. Monthly returns are formed using the open interest weighted price across all strike prices from the last trading day of each month. I do not delta-hedge options due to the costly nature, from a portfolio formation perspective, of delta-hedging.

It is well-known that options data contain observation errors at a rate much larger than equity data (Todorov (2019), Andersen et al. (2021)). Common causes of these errors include the true option value lying within the bid-ask spread (minimum tick sizes for options are usually 5 to 10 times the size of stock minimum tick sizes, and stock price is an input into option pricing formulas), liquidity provisions offered by market makers and shifts in option positions which effect only a small range of strike prices. To mitigate the impact of these observation errors I apply a number of filters to the data before calculating monthly returns.

The filters include a bid-ask spread larger than the bid-ask midpoint (Gharghori et al., 2017), zero open interest (Goyal and Saretto, 2009), option price less than \$0.10 and winsorizing all variables at the 0.1% level (Muravyev and Pearson, 2020). Because I do not want options to expire during the portfolio holding period, I also require options to have a time to execution of between one and two years and an implied volatility between 0 and 2. To further mitigate any error in reported options pricing I winsorize the generated monthly returns at the 0.1% level as well. This has the added benefit of ensuring my results are not driven by a single outlier return. From January 1996 to December 2017 WRDS has already pre-processed the data to generate variables such as price — which is not available from OptionMetrics — and pre-filtered on implied volatility and time to execution. I extend the series to December 2020 using raw OptionMetrics data. Table 5 summarizes the option data cleaning process.

Table 5: Option Data Filtering Procedure

Filter	Observations	Deleted	% Deleted
Panel (A) Pre-filtered by WRDS (1996 - 2017): implied volatility $\in$ (0,2) and TTE $\in$ (366,730)			
Total Daily Closing Prices	**78,148,602**		
Call options only	39,994,641	38,153,961	48.82%
Spread larger than price	38,619,531	1,375,110	1.76%
Price less than \$0.1	38,498,683	120,848	0.15%
Zero open interest	38,498,683	0	0%
Aggregate across strikes and maturities to daily and keep last trading day	179,013		
Panel (B) OptionMetrics Raw Data (2018 - 2020)			
Total Daily Closing Prices	**779,601,687**		
Implied volatility $\in$ (0,2)	632,266,392	147,335,295	18.90%
Time to execution $\in$ (366,730)	45,179,467	587,086,926	75.31%
Call options only	22,057,664	23,121,803	2.97%
Spread larger than price	20,393,129	1,664,535	0.21%
Price less than \$0.1	20,342,661	50,468	0.01%
Zero open interest	16,565,647	3,777,014	0.48%
Aggregate across strikes and maturities to daily and keep last trading day	36,108		
Final firm-month observations	**215,121**		

Unlike the equity and bond data, delisting or default returns are not a concern for options. During the period 1996 to 2020, the options of companies with stock that was delisted for any reason reach a price of zero (or a value that is not different from zero by any economically or statistically relevant amount) well-before the delisting event. This finding is echoed in the literature as no paper I am aware of attempts to add delisting returns to options series.

SSD_c is generated using single-name CDS price data from Bloomberg for the period October 2001 through December 2020. CDS contracts exist for a much smaller universe of firms compared to equity and option securities and are generally less liquid. I therefore limit CDS data to firms belonging to the S&P 500 index in May 2021 to ensure sufficient CDS contracts trade each month to generate characteristic sorted portfolios. Following Meine et al. (2016) I use single name CDS contracts with a 5-year term structure because they are the most liquid. Despite being the most frequently traded, only 237 of the 500 firms belonging to the S&P 500 have actively traded credit default swaps.

To protect against possible errors in CDS prices I also impose the following filters before calculating monthly CDS returns which, like option returns, are calculated as the percent change in price on the last trading day of each month. First I drop all prices above 10,000. CDS prices are denominated in basis points of the debt position the CDS is insuring. Therefore a price above 10,000 would imply it costs more to insure the debt than the face value of the debt. This filter applies to only seven observations, all American Airlines Group during the market downturn in 2009 and the Covid-19 pandemic in 2020. I then winsorize price at the 0.1% level before generating monthly returns. Lastly I winsorize the return series at the 0.1% level to ensure the results are not driven by a single outlier. Table 6 summarizes the CDS data cleaning process.

Table 6: CDS Data Filtering Procedure

Filter	Observations	Deleted	% Deleted
Total Daily Prices - Single Name 5-Year Term Structure	**855,342**		
Prices above 10,000	855,335	7	0.00%
Keep last trading day of month, require consecutive months	37,907		
Final firm-month observations	**37,907**		

4 Bankruptcy Risk Factor

This section presents the returns generated by SSD_i (for $i \in (e, b, o, c)$) in the equity, bond, option and CDS markets. For each of the four factors I present evidence of the statistical significance and economic benefits to an investor.

4.1 Equity Factor

To motivate the formation of the bankruptcy risk factor, I first present decile portfolio sorts on bankruptcy risk as measured by Altman's Z-score, the WLS update of the Z-score and the random forest probability of bankruptcy. Below the portfolio returns are the associated t-statistic as well as the mean probability of bankruptcy within that portfolio.

Looking first at returns, the pattern displayed in the extreme decile long-short portfolio (column (10)-(1)) is as we would expect given my hypothesis regarding the evolution of the risks facing firms over time. The portfolios sorted by Altman's Z-score exhibit a much stronger pattern when the sample begins in 1962 compared to when the sample begins in 1980 demonstrating the Z-score was more relevant for predicting returns in the 1960s and 70s than it is today. Similarly, the WLS update of the coefficients associated with the five accounting ratios used by Altman (1968) lead to better defined portfolio sorts, but the relationship is still stronger in the 1960s and 70s than it is today. This implies both the weights and inputs used to construct the Z-score are obsolete in the context of predicting equity returns.

On the other hand, sorts based on the random forest estimated probability of bankruptcy lead to better discrimination between high and low bankruptcy risk firms — particularly during the more recent time period. This provides evidence for the benefits of using contemporary risk factors to measure bankruptcy probability.

Table 7: Equity Decile Mean Returns and Bankruptcy Risk

Sort	Distress (1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	Safe (10)	(10)-(1)
AltZ 1962	1.13%	1.02%	0.89%	0.93%	0.99%	1.02%	1.01%	0.90%	0.94%	0.90%	-0.23%
t	(5.67)	(5.99)	(5.26)	(5.12)	(5.57)	(5.79)	(5.68)	(5.15)	(5.41)	(4.28)	(1.39)
$\bar{Z}$	2.11	4.12	5.46	6.87	8.56	10.76	13.82	18.64	28.43	169.64	
AltZ 1980	1.12%	1.14%	1.02%	1.01%	1.11%	1.11%	1.10%	0.99%	1.10%	1.03%	-0.09%
t	(4.93)	(5.94)	(5.27)	(4.82)	(5.33)	(5.52)	(5.38)	(4.84)	(5.56)	(4.06)	(0.44)
$\bar{Z}$	2.07	4.20	5.65	7.16	8.23	11.26	14.46	19.49	29.36	185.79	
WLS 1962	0.86%	0.84%	0.96%	0.90%	0.94%	0.97%	0.98%	1.03%	1.07%	1.10%	0.24%
t	(5.25)	(5.09)	(5.58)	(4.90)	(4.81)	(4.89)	(4.75)	(4.90)	(4.99)	(5.62)	(1.41)
$\bar{Z}_{WLS}$	0.58	0.30	0.23	0.19	0.16	0.13	0.10	0.06	0.02	-0.09	
WLS 1980	1.01%	0.91%	1.09%	1.00%	1.10%	1.10%	1.09%	1.23%	1.22%	1.18%	0.17%
t	(5.01)	(4.76)	(5.13)	(4.53)	(4.74)	(4.71)	(4.38)	(4.82)	(4.76)	(5.22)	(1.31)
$\bar{Z}_{WLS}$	0.59	0.30	0.23	0.19	0.15	0.12	0.09	0.06	0.02	-0.09	
RF 1962	0.97%	0.86%	0.86%	0.95%	1.10%	1.03%	0.86%	0.88%	1.00%	1.08%	0.11%
t	(4.05)	(4.19)	(4.37)	(5.50)	(6.33)	(7.71)	(5.17)	(5.53)	(5.92)	(5.28)	(0.78)
$\bar{Z}_{RF}$	0.79	0.58	0.39	0.23	0.13	0.07	0.04	0.03	0.02	0.01	
RF 1980	1.02%	0.88%	0.90%	0.94%	1.12%	1.12%	1.00%	0.96%	1.04%	1.20%	0.18%
t	(3.60)	(3.45)	(3.72)	(4.61)	(5.47)	(5.91)	(4.94)	(4.99)	(5.26)	(4.64)	(1.12)
$\bar{Z}_{RF}$	0.83	0.64	0.43	0.26	0.14	0.08	0.04	0.03	0.02	0.01	

Note: This table presents mean returns for various portfolios formed by sorting firms based on the original Altman Z-score (AltZ), a weighted least squares re-estimation of the weights used to form the Z-score (WLS) and a random forest (RF). Rows 1962 calculate values using the sample from 1962 to 2019. Rows 1980 calculate values using the sample from 1980 to 2019. t-statistics are in parentheses. $\bar{Z}$ represents the mean of the bankruptcy risk measure used in the respective row for each portfolio. Columns Distress (1) to (10)-(1) show mean returns for single sorts on bankruptcy risk.

Despite the better defined division between high and low bankruptcy risk firms produced by the random forest, the extreme decile long-short portfolio is not statistically significant for any of the bankruptcy risk models. This is consistent with Dichev (1998) who shows using both the Altman Z and Ohlson O measures of bankruptcy risk that portfolios sorted on bankruptcy risk do not necessarily form monotonic return patterns. There are two possible explanations for this. The first is that investors — familiar with the theoretical relationship between risk and return — demand risky stocks due to a misunderstanding surrounding empirical return patterns.

The second is related to the mean bankruptcy probability in each decile portfolio. Table 7 shows the bankruptcy probability for the five safest deciles are largely similar — even the mean bankruptcy probability for decile 5 is only six percentage points higher than decile 6. It is not surprising then that the returns for these decile portfolios are so similar given the similarity in bankruptcy risk. However, the fact the highest return decile is always among the three safest deciles does provide further evidence in support of the residual income model of Asness, Frazzini and Pedersen (2019).

The equity bankruptcy risk factor is formed by sorting firms independently by size and my random forest estimated probability of bankruptcy. Sorts are done in June using accounting and market equity information released in December of the previous year. Thus the assumption is made that firm risk information has been fully disseminated and absorbed by investors within six months of its release. Motivated by the observation that bankruptcy risk is very similar among deciles 6 through 10, and bankruptcy risk is unreasonably high in decile 1, the bankruptcy risk factor is formed by going long an equal weighted average of the five safest deciles within both the large and small size groups, and going short an equal weighted average of decile two within the large and small size groups. Shorting decile two instead of decile one avoids the firms with the most uncertain behavior — decile one consistently has the lowest standard errors among the decile portfolios — removing unnecessary risk from the factor. Dichev (1998) also forms portfolios which have unequal numbers of portfolios in the long and short legs.

Dichev (1998) also demonstrates that distress risk is unrelated to the size anomaly. If this is still true, then in small firms should see higher returns than large firms, and safe firms should see higher returns than distressed firms. This means safe small firms should have the highest returns, while large distressed firms should have the smallest. Table 8 confirms these return patterns and presents the average returns and maximum drawdowns of the bankruptcy risk factor termed SSD_e for "safe subtracting distressed"[9].

Consistent with Table 7, the same pattern emerges as we transition from the factor constructed on Altman's Z-score to the factor constructed using my random forest bankruptcy probability. Raw and risk adjusted returns are highest and maximum drawdowns are lowest for the random forest model. This provides yet further evidence supporting my hypothesis

[9]A better name may be "safe minus distressed", but the acronym — SMD — is too similar to the Fama-French size factor SMB so I avoid it to avoid confusion.

that it is important to not only update the weights associated with a bankruptcy prediction model, but also to identify the appropriate contemporary risk factors facing firms.

Table 8: Bankruptcy Risk Equity Factor: 1980 - 2019

			Original Altman			
	High (2)	Med (3-5)	Low (6-10)	SSD_e	t(SSD_e)	Max Drawdown
Small	1.23%	1.20%	1.08%	0.10%	0.73	55.0%
Large	0.98%	0.90%	1.09%			
			WLS Re-estimation			
	High (2)	Med (3-5)	Low (6-10)	SSD_e	t(SSD_e)	Max Drawdown
Small	1.00%	1.03%	1.19%	0.18%	1.73	31.9%
Large	0.89%	0.92%	0.96%			
			Random Forest			
	High (2)	Med (3-5)	Low (6-10)	SSD_e	t(SSD_e)	Max Drawdown
Small	1.02%	1.13%	1.23%	0.23%	2.29	29.4%
Large	0.87%	0.88%	0.96%			

Note: This table presents mean returns for the components of the equity bankruptcy risk factor SSD_e as well as the factor itself. SSD_e is formed by sorting firms in June independently by size and my random forest estimated probability of bankruptcy. The factor is long an equal weighted average of the five safest deciles of large firms and the five safest deciles of small firms, and short an equal weighted average of the second riskiest decile of large firms and the second riskiest decile of small firms. The numbers in parentheses in the columns represent the included deciles from Table 7. Max Drawdown represents the maximum peak to trough decrease from 1980 to 2019 and does not represent any specific period length.

Although the average monthly returns of SSD_e are not huge, they are of similar size to other equity factors in the literature. Alwathainani (2009) documented a monthly return of 0.21% for his factor sorted on earnings consistency which is validated by Chen and Zimmermann (*forthcoming*). Hirschleifer et al. (2013) documented a monthly return of 0.26% for their factor based on patents scaled by R&D expenditures, although Chen and Zimmerman find only 0.21% returns on the same factor. Hou and Robinson (2006) documented monthly returns of 0.26% for their factor sorted on industry concentration, although Chen and Zimmermann find only 0.21% returns on the same factor.

Additionally, the information contained in SSD_e is independent of existing factors. SSD_e has positive alpha — which can be interpreted as risk-adjusted returns since the influence of other types of risk have been accounted for by the right-hand-side factors — when regressed on the market or the Fama-French five-factors. Although it is commonplace (as it should be) to test new factors against influential asset pricing models such as the Fama-French five factor model, it is still important to test new factors against the CAPM by itself because alphas can move from insignificant to significant as more factors are added (Jensen, Kelly and Pedersen, 2021). A successful new factors should achieve a statistically significant alpha against both.

The coefficient on SSD_e is also statistically significant in cross-sectional regressions when it is included along with 15 other existing factors selected using the two-step lasso procedure

developed by Feng, Giglio and Xiu (2020). The procedure of Feng, Giglio and Xiu (2020) was developed as a way to provide evidence of the incremental explanatory ability of a proposed factor for the cross-section of returns in the face of the myriad existing factors in the literature — often called the "factor zoo". I use the 135 factors in Chen and Zimmermann (*forthcoming*) which form a balanced panel from 1964 to 2019 as a proxy for the factor zoo. The 15 factors selected by this procedure of summarized in Appendix D, while the SSD_e regressions are summarized in Table 9.

Table 9: SSD_e Risk-Adjusted Monthly Returns

	CAPM	FF5	Zoo
SSD_e	0.25%	0.23%	5.34
t(SSD_e)	(3.39)***	(2.60)***	(5.93)***

Note: This table presents the statistical significance of SSD_e when regressed on other popular factors. For the columns CAPM (regression on the capital asset pricing model) and FF5 (regression on the five Fama and French factors) the table reports the regression alphas and corresponding t-statistics. The column Zoo reports the coefficient and t-statistic on SSD_e when it is included in cross-sectional regressions along with 15 other factors chosen using the two-step method proposed by Feng, Giglio and Xiu (2020). *** indicates statistical significance at the 1% level.

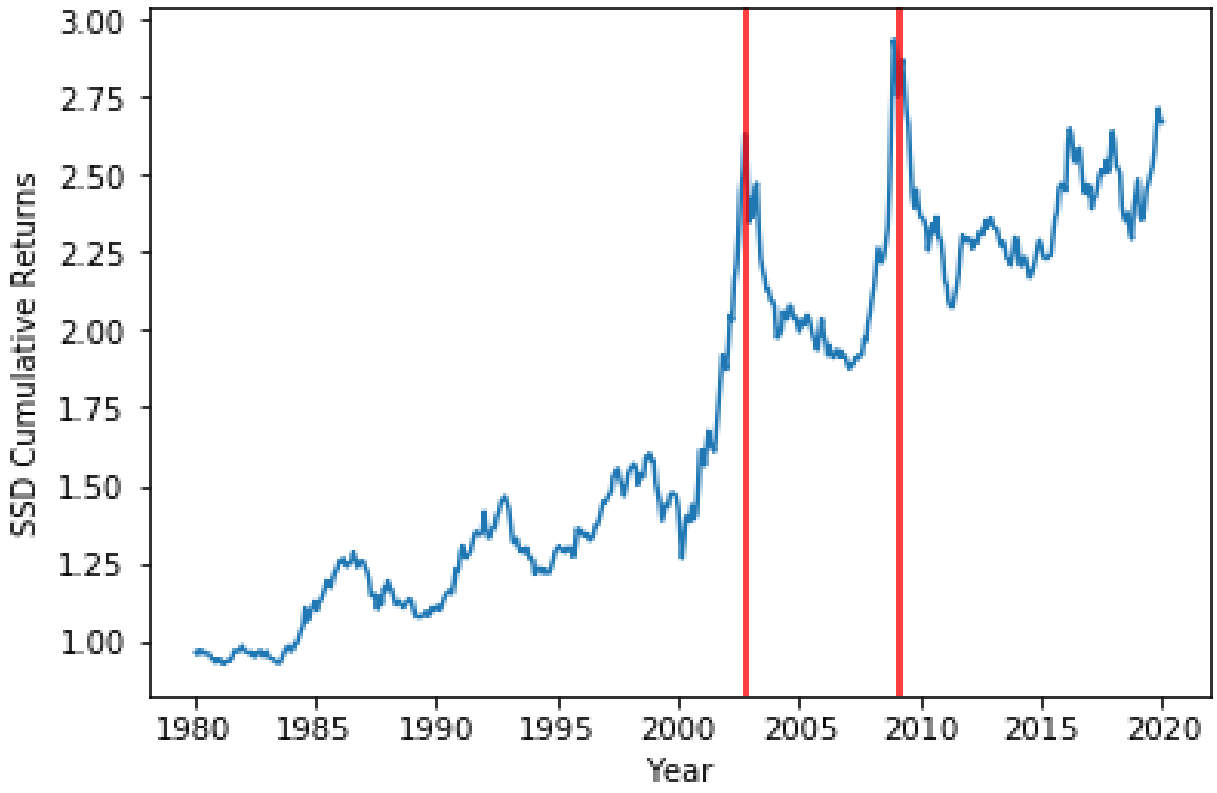
The risk-adjusted returns of SSD_e remain fairly close to their raw value of 0.23% per month when regressed on the CAPM and Fama-French five-factor models. The alpha when regressed on the CAPM is 0.25% per month and 0.23% per month when regressed on the Fama-French five-factor model. The coefficient 5.34 for the Zoo column has a related, but different interpretation. It is the stochastic discount factor loading on SSD_e controlling for the other 15 selected factors when explaining the cross-section of returns. All three coefficients are highly statistically significant.

The statistically significant raw and risk-adjusted returns, coupled with the statistically significant stochastic discount factor loading on SSD_e in the presence of the factor zoo provides strong evidence that SSD_e contains information which is orthogonal to existing factors. It appears the five ratios selected by the random forest, though largely absent individually in the literature as characteristics used to form portfolios, can provide important information when combined in the right way.

Figure 3 shows an alternative presentation of the information in Table 8 — the cumulative returns to \$1 invested in SSD_e in 1980. Despite the relatively slow growth of SSD_e over time (0.23% per month), there are some desirable features of this returns series. First, SSD_e is a leading indicator of market recovery after crashes. The two vertical lines in Figure 3 indicate market low points during the dot com bubble and the Great Recession. SSD_e sees huge increases in value well before the market begins to recover — providing evidence for the flight-to-safety phenomenon.

Second, the maximum drawdown of SSD_e is 29.4% compared to 59% for the S&P 500 index (which occurred during the Great Recession) and 58% for the momentum factor (also

Figure 3: Cumulative Returns to SSD_e Investment Strategy: 1980 - 2019



Note: This figure presents the cumulative returns to \$1 invested in SSD_e in 1980. The vertical lines indicate the lowest point of the market downturns during the "dot com" bubble (September 2002) and Great Recession (February 2009).

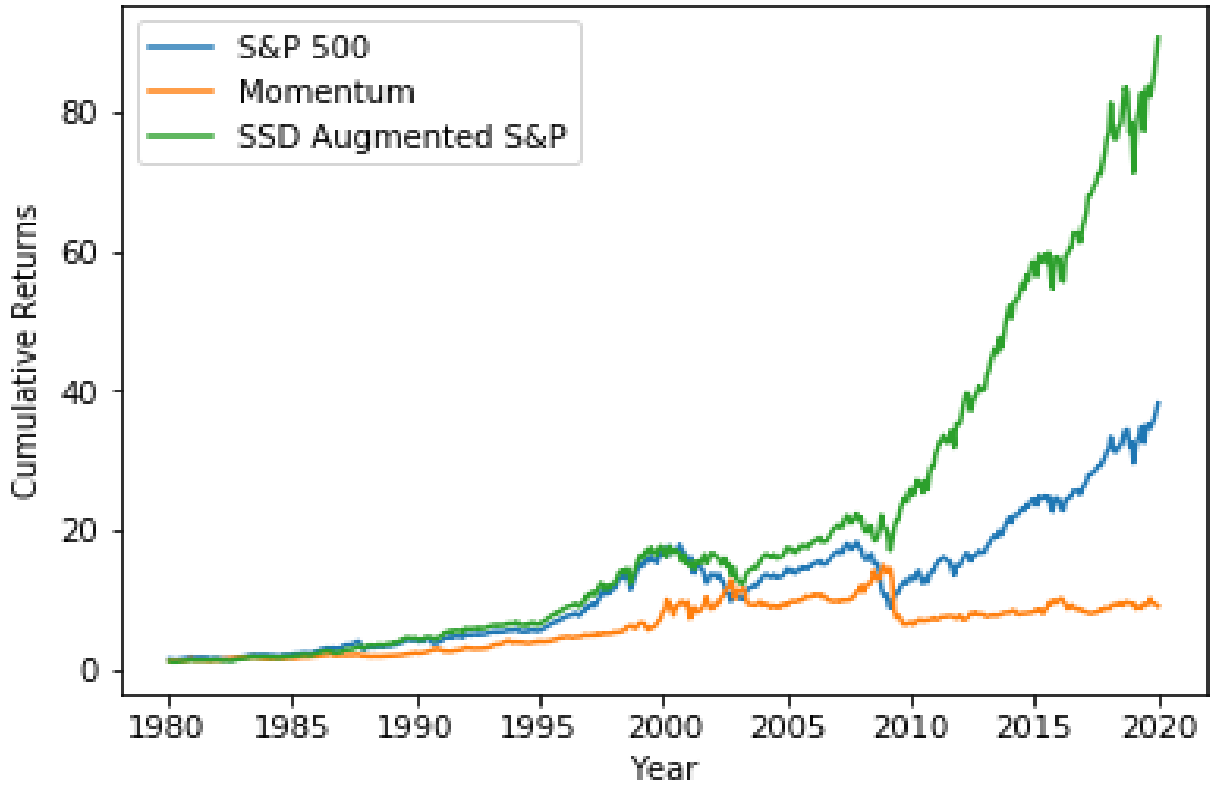
during the Great Recession). This makes SSD_e in its particularly relevant for leverage investors, as these investors could apply more than three-times leverage in SSD_e whereas even two-times leverage would completely kill an investors portfolio if investing in the S&P 500 index or momentum.

In Appendix E I present a comparison of SSD_e formed using my random forest measure of bankruptcy risk to portfolio sorts based on Altman's Z-Score and the WLS update of the Z-Score. Not only are the cumulative returns lower for the two other models, but neither are leading indicators of market recovery. This removes one of the main benefits of SSD_e as demonstrated in Figure 4.

Taking advantage of the leading indicator aspect of SSD_e, an investor can augment — for example — a buy and hold strategy as follows. Once investors observe large negative market returns — -5% in a month — the following month invest in SSD_e until index returns rise above -5%. Implementing this strategy results in an average monthly return of 1.02% — higher than the raw buy and hold return of 0.80% per month and momentum return of 0.56% per month.

It is clear from Figure 4 that the benefits of this hybrid strategy come from significantly shortened period of large negative returns. By moving investments into SSD_e after observing large decreases in market returns, investors both avoid continued negative buy and hold returns, and are fully invested in the safe stocks before other investors drive up prices. This strategy results in an annualized Sharpe ratio of 0.90, higher than the annualized Sharpe ratios of the buy and hold return (0.65) and momentum strategy (0.44) during the same period.

Figure 4: Cumulative Returns to SSD_e Augmented Buy and Hold Return: 1980 - 2019



4.2 Bond Factor

As with SSD_e, I motivate the formation of the bond bankruptcy risk factor by first presenting decile portfolios of bonds sorted on the random forest estimated probability of bankruptcy. Since bond trade data is only available through enhanced TRACE beginning in 2002 I cannot compare portfolios formed in the 1960s and 1980s as I did for the equity decile portfolios. Additionally, the relative return distribution for the portfolios sorted by Altman's Z-score, the WLS update of the Z-score coefficients and the random forest estimated bankruptcy probability are very similar to the pattern observed in the equity market. For clarity of exposition I therefore only present results for the portfolios sorted on the random forest probability of bankruptcy. These results are displayed in Table 10.

Table 10: Bond Decile Mean Returns and Bankruptcy Risk

Sort	Distress (1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	Safe (10)	(1)-(10)
RF 2002	0.75%	0.69%	0.73%	0.65%	0.61%	0.58%	0.57%	0.56%	0.58%	0.58%	0.17%
t	(4.35)	(4.70)	(5.62)	(5.19)	(5.48)	(5.08)	(5.06)	(5.11)	(5.42)	(5.33)	(1.50)
$\bar{Z}_{RF}$	0.79	0.56	0.39	0.26	0.17	0.11	0.06	0.04	0.02	0.01	

Note: This table presents mean returns for various portfolios formed by sorting firms based on my random forest produced probability of bankruptcy. The row RF 2002 presents mean returns for each decile. t-statistics are in parentheses. $\bar{Z}$ represents the mean of the bankruptcy risk measure used for each portfolio.

Unlike equity returns, bond returns are not driven by changes in the clean price of bonds, but rather by accrued interest and coupon payments. It therefore makes sense that we do not observe the same high returns to bonds with low bankruptcy risk, driven by the perceived undervaluation of safer assets as argued by Asness, Frazzini and Pedersen (2019). Bonds issued by companies with higher default risk — which is related to my measure of bankruptcy probability — must generally pay higher interest rates. This leads to higher returns on bonds with high bankruptcy (default) risk. However, Table 10 still highlights the increased uncertainty surrounding returns to the riskiest bonds. The relatively low t-statistic for the most distressed decile of bond returns implies highly distressed firms default at higher rates than safer firms.

The pattern of returns across the cross-section of bonds is very similar to the cross-section of equities, albeit with returns increasing in the opposite direction. Specifically, while returns are not monotonically decreasing in safety, there is a clear pattern of decreasing returns until the probability of bankruptcy reaches approximately 10%. After that point there is materially no difference in returns among the remainder of the decile portfolios resulting in a difference in extreme decile returns which is not statistically significant. There is one more element of note in Table 10 before I motivate formation of the bankruptcy risk bond factor. The three riskiest deciles have lower mean levels of bankruptcy risk than their equity counterparts. This makes sense, since these firms would have to pay the highest interest rates on bond issues, raising capital via the stock market may be a cheaper alternative.

Table 11: Bankruptcy Risk Bond Factor: 2002 - 2020

	High (1-3)	Med (4-5)	Low (6-10)	SSD_e	t(SSD_e)	Max Drawdown	SR
Small	0.74%	0.65%	0.59%	0.14%	2.77	6.3%	0.69
Large	0.69%	0.56%	0.55%				

Note: This table presents mean returns for the components of the bond bankruptcy risk factor SSD_b as well as the factor itself. SSD_b is formed by sorting firms in June independently on size and my random forest estimated probability of bankruptcy. The factor is formed by going long an equal weighted average of the three riskiest deciles for large firms and the three riskiest deciles for small firms, and going short an equal weighted average of the five safest deciles for large firms and the give safest deciles for small firms. The numbers in parentheses in the columns represent the included deciles from Table 10. Max Drawdown represents the maximum peak to trough decrease from 2002 to 2020 and does not represent any specific period length.

SSD_b is formed in the same spirit as SSD_e. That is, in June firms are sorted independently by size and my random forest estimated probability of bankruptcy. Motivated by the

information in Table 10, SSD_b is formed by going long an equal weighted average of the three riskiest deciles within both the large and small size groups, and going short an equal weighted average of the five safest deciles within the large and small size groups. The only difference between the formation of SSD_b and SSD_e — that the three riskiest deciles are included in the bond factor instead of just decile 2 — reflects the fact that bonds have both built-in compensation for extra risk in terms of higher interest rates, and lower probability of bankruptcy compared to the equity portfolios for the three riskiest deciles.

Crawford et al. (2019) demonstrates that size is a priced risk factor in the bond market where smaller firms — being riskier — require higher interest rates and therefore have higher returns. I should therefore find that smaller firms have higher returns than larger firms for a given level of bankruptcy risk, and within size groups higher bankruptcy risk should result in higher returns. Table 11 shows this is indeed the case. The result is the bond factor, termed SSD_b (where the b is for "bond") earns a statistically significant 0.15% return per month.

Table 12: SSD_b Risk-Adjusted Monthly Returns

	Market	Bond6	FF6	FF-Bond12
SSD_b	0.11%	0.08%	0.08%	0.08%
t-statistic	(2.27)**	(2.08)**	(2.08)**	(2.16)**

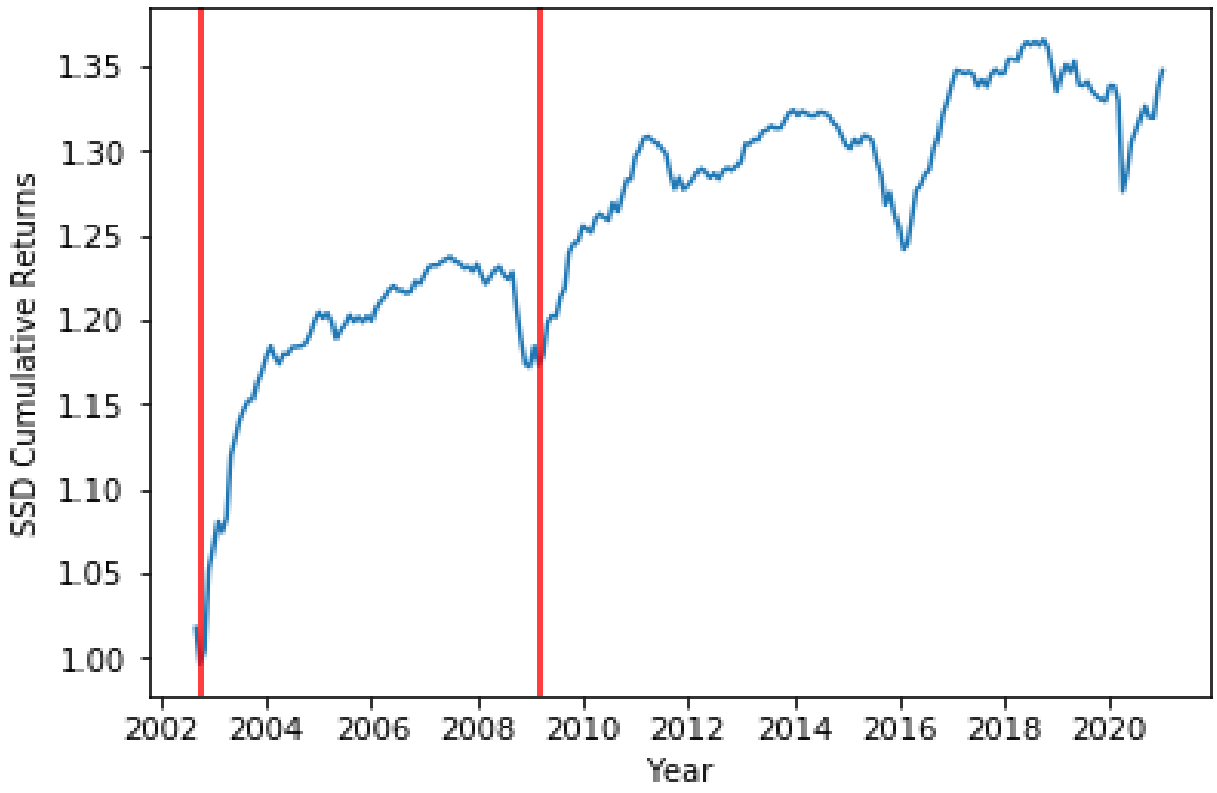
Note: This table presents the statistical significance of SSD_b when regressed on other popular factors. *Market* indicates a regression on the aggregate bond market return. *Bond6* indicates a regression on six standard bond market factors. *FF6* indicates a regression on the Fama-French six equity factors (including momentum). *FF-Bond12* indicates a regression on all 12 previous used equity and bond factors. ** indicates statistical significance at the 5% level.

Like SSD_e this is not an enormous monthly return. However, it is independent of existing bond and equity factors in the literature. Following Bai, Bali and Wen (2019), I regress SSD_b on both equity and bond factors to obtain alpha, or the risk-adjusted returns. Controlling for the aggregate bond market return in excess of the risk-free rate, SSD_b earns a positive 0.11% per month. This decreases to a 0.08% risk-adjusted monthly return when five other standard bond factors are added. These factors are bond momentum (Jostova et al., 2013), two measure of liquidity (Bao, Pan and Wang, 2011; and Bai, Bali and Wen, 2019)[10], and default and term factors (Elton et al., 1995).

SSD_b maintains a risk-adjusted return of 0.08% per month when regressed on the Fama-French six equity factors and when both equity and bond factors are included in the regressions. That SSD_b consistently maintains statistically significant risk-adjusted returns while moving from a univariate regression on the excess bond market return to a 12 factor model provides strong evidence for the independence of the information it contains.

[10]I include two liquidity factors because Bai, Bali and Wen (2019) include one liquidity factor in their spanning regressions but they also introduce a new liquidity factor. Since neither liquidity factor subsumes the other in a univariate spanning regression I chose to include both.

Figure 5: Cumulative Returns to SSD_b Investment Strategy: 2002 - 2020



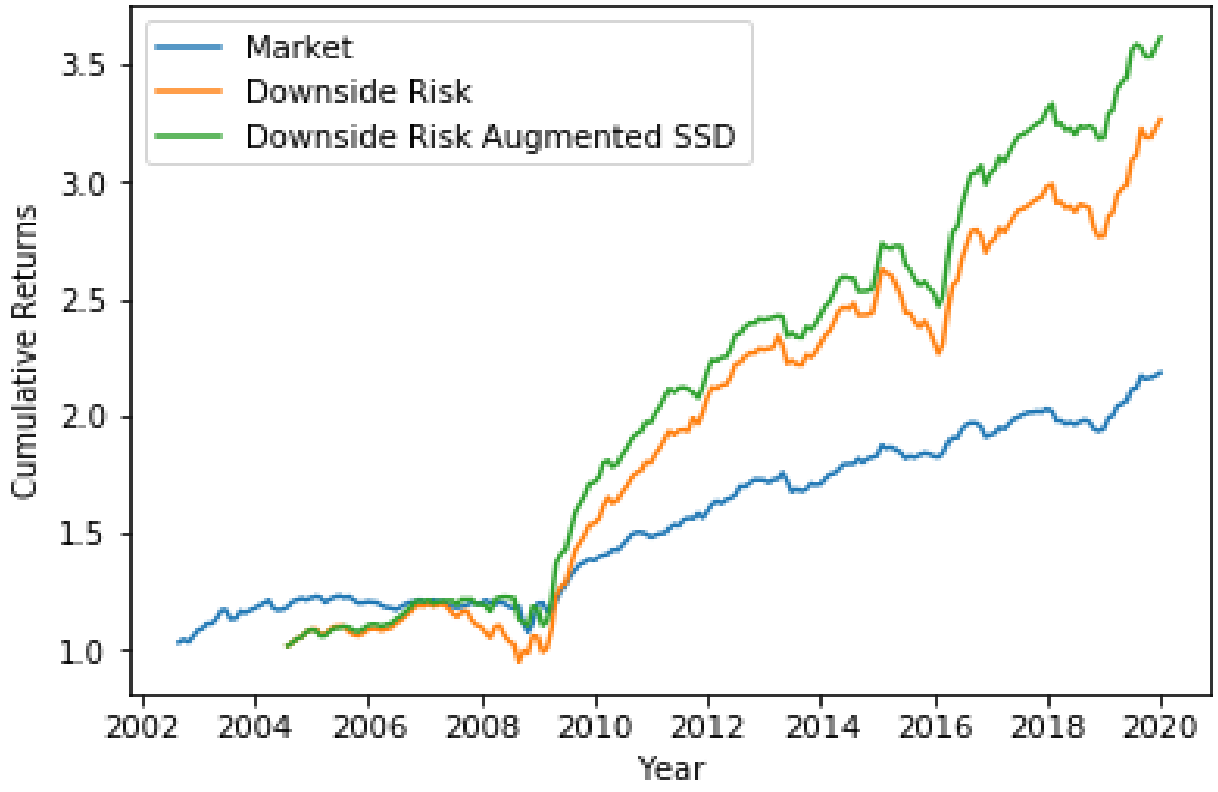
Note: This figure presents the cumulative returns to \$1 invested in SSD_b in 2002. The vertical lines indicate the lowest point of the market downturns during the "dot com" bubble (September 2002) and Great Recession (February 2009).

Figure 5 provides an alternative presentation of the information contained in Table 11. The source of the dips in 2009 and 2020 are obvious — the Great Recession and the onset of the Covid-19 pandemic. The source of the dip in 2016-2017 is less obvious, although it is common in existing bond factors. As with SSD_e, I present a comparison of SSD_b with factors generated by sorting on Altman's Z-score and the WLS update of the Z-score coefficients in Appendix E.

While the primary benefits of SSD_e are from its role as a leading indicator of market recovery in the equity space, Figure 5 shows that SSD_b clearly does not lead recoveries in the bond market. The major benefits of SSD_b come from the small maximum drawdowns during periods of systemic distress. These small drawdowns — the largest of which is only 6.3% compared to 11.1% for the bond market and 20.9% for downside risk — and low levels of overall volatility are such that SSD_b has a Sharpe ratio of 0.69 despite averaging only a 0.15% monthly return. Motivated by the relatively small decreases during periods of systemic distress, a similar strategy to the one proposed in section 4.1 is proposed here — using SSD_b to augment an existing investment strategy.

Specifically, beginning with the downside risk factor (DRF) of Bai, Bali and Wen (2019), hold the DRF portfolio until observing a monthly decrease of -1%. In the subsequent month transfer the investment to SSD_b until returns on DRF rise above -1%, then return the investment to DRF the following month. In the same way the equity hybrid strategy took advantage of SSD_e as a leading indicator, this strategy takes advantage of SSD_b as a safety net during periods of distress as the bottom of the downturn is less severe than that of other strategies. Implementing this strategy results in monthly returns of 0.71% per month compared to 0.66% for DRF and 0.38% for the market. The benefits of this strategy can also be seen graphically in Figure 6 and in the associated Sharpe ratio which is 1.20 for this bond hybrid strategy, compared to 1.02 for DRF and 0.98 for the aggregate bond market. Importantly, any bond factor strategy can be substituted for DRF. DRF was chosen for illustrative purposes because of its large raw return and high Sharpe ratio.

Figure 6: Cumulative Returns to SSD_b Augmented Investment: 2002 - 2020



4.3 Options Factor

In the same way I motivated the formation of SSD_e and SSD_b by first looking at decile portfolios sorts, I motivate the options bankruptcy risk factor by first examining the average returns and bankruptcy risk of the cross-section of call options. It is important to redo this exercise for the cross-section of call options because there are fewer firms with actively traded options than there are with actively traded stocks (there are 21,803 firms with returns data

appearing in CRSP at least one month during the period 1996-2020, but only 2,706 firms with eligible options returns). This could impact the mean bankruptcy risk in each decile of the cross-section, impacting how the options bankruptcy factor is formed.

Since the Ivy OptionMetrics database only has options data beginning in 1996 I cannot compare portfolios formed in the 1960s and 1980s like I did for the equity decile returns. Additionally, the relative return distribution for the portfolios sorted by Altman's Z-score, the WLS update of the Z-Score coefficients, and the random forest probability of bankruptcy are very similar to the pattern observed for the equity factor. Therefore for clarity of exposition I only display results for the portfolios sorted on the random forest bankruptcy probability. These results are displayed in 13. The equivalent results for portfolios formed on Altman's Z and the WLS models are available upon request.

Table 13: Option Decile Mean Returns and Bankruptcy Risk

Sort	Distress (1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	Safe (10)	(1)-(10)
RF 1996	9.20%	9.52%	8.77%	8.08%	7.93%	6.94%	7.57%	7.83%	7.93%	7.58%	1.62%
t	(6.44)	(6.18)	(6.09)	(6.48)	(5.65)	(5.78)	(5.94)	(6.17)	(5.89)	(5.11)	(2.42)
$\bar{Z}_{RF}$	0.86	0.70	0.50	0.31	0.17	0.09	0.05	0.03	0.02	0.01	

Note: This table presents mean returns for various portfolios formed by sorting firms based on my random forest produced probability of bankruptcy. The row RF 1996 presents mean returns for each decile. t-statistics are in parentheses. $\bar{Z}$ represents the mean of the bankruptcy risk measure used for each portfolio.

Call option returns like equity returns are not monotonic. Despite this fact, the pattern in the cross-section of call option returns is stronger than for the equity market reflecting the increased sensitivity of options investors to subtle differences in risks facing underlying asset returns. Despite the stronger pattern in mean returns, the pattern of mean bankruptcy risk for each portfolio is largely similar to the equity portfolios. Deciles 6 through 10 have very similar probabilities of bankruptcy and decile 1 has a very high probability of bankruptcy. This pattern suggests portfolio formation consistent with SSD_e.

Table 14: Bankruptcy Risk Option Factor: 1996 - 2020

	High (9)	Mid (6-8)	Low (1-5)	SSD_o	t(SSD_o)	Max Drawdown	SR
SSD_o	9.52%	8.32%	7.55%	1.97%	3.01	46.8%	0.61

Note: This table presents information related to the option bankruptcy risk factor. SSD_o is formed by sorting firms in June on my random forest estimated probability of bankruptcy into decile portfolios and going long call options in the second riskiest decile and going short on an equal weighted average of the five safest deciles. SSD_o is the mean monthly returns from the long-short bankruptcy risk factor, t(SSD_o) is the monthly t-statistic associated with the factor. Max drawdown is the maximum peak-to-trough decrease over the full 22 year sample without regard to a specific time period. SR is the annualized Sharpe ratio. The number is parentheses ((6-8) for example) represent the component decile portfolios from Table 13.

The options factor is formed on a univariate sort of the random forest estimated probability of bankruptcy. Specifically, each year firms are sorted in June on my bankruptcy risk probability generated with accounting and market equity information released in December of the previous year. The options factor — titled SSD_o where the subscript is for "options" — is formed by going long on call options belonging to the second decile of Table 13 (avoiding the

most distressed firms which have an unreasonably high probability of bankruptcy) and going short (shorting call options is very similar to, but not exactly the same as, going long on a put option) on an equal weighted average of call options for the five safest deciles of Table 13.

Because SSD_o is formed using a univariate sort instead of an independent bivariate sort on both bankruptcy risk and size, there could be a concern that the returns to SSD_o are driven by size instead of bankruptcy risk since size is such a strong predictor of bankruptcy. I demonstrate in Table 15 that SSD_o is independent of existing option factors — including size. However, to further quell fears, I also note that portfolio turnover is much higher for the bankruptcy risk deciles, than size deciles. The largest firms — as proxied by S&P 500 membership have an annual turnover of only 2.3%, while the safest bankruptcy decile has an annual turnover of 28.1%. This makes it very unlikely returns to SSD_o are driven by the size anomaly.

Like the cross-section of bond returns, the cross-section of call option returns are increasing in the opposite direction of equity returns. The reversed direction of the call option sort can be explained by two risk related concepts — risk aversion and prudence (Kimball, 1990). The variance risk premium of equity options is negative (Christoffersen et al. (2018)), implying the expected variance is higher under the risk neutral distribution than under the physical distribution, driving returns for out-of-the-money (OTM) call options. Additionally, as bankruptcy risk increases (going from column (10) to column (1) in Table 13) investors with negative prudence drive up the price of OTM call options as speculators buy them to use as levered bets.

Table 15: SSD_o Risk-Adjusted Monthly Returns

	Z	Option6	FF6
SSD_o	1.33%	2.14%	2.92%
t(SSD_o)	(2.17)**	(2.76)***	(4.71)***

Note: This table presents the statistical significance of SSD_o when regressed on other popular factors. Column *Z* regresses SSD_o on an option factor generated by sorting firms on Altman's Z-score. Column *Option6* regresses SSD_o on six option factors (implied volatility, illiquidity, size, the difference between implied and realized volatility, idiosyncratic volatility and book-to-market). Column *FF6* regresses SSD_o on the six Fama-French equity factors. t-statistics are in parentheses. ** indicates significance at the 5% level, *** indicates statistical significance at the 1% level.

The values in Table 14 — the largest among the four factors introduced in this paper — compare favorably to other popular investment vehicles. Again using as example the momentum factor and the buy and hold S&P 500 index return, a mean monthly return of 1.97% over the period 1996 to 2020 is larger than the 0.41% and 0.67% mean monthly returns of momentum and buy and hold strategies, respectively.

While the raw returns of SSD_o are impressive, following Cao et al. (*forthcoming*) and Horenstein et al. (2020) I regress SSD_o on existing equity and option factors to verify the information SSD_o contains is independent of existing factors in the literature. Unlike

the equity and bond literatures, there are very few anomaly based option pricing models from which to pick factors for spanning regressions. I therefore choose to regress SSD_o on the Altman Z option factor, a six option factor model (implied volatility, illiquidity, size, the difference between implied volatility and realized volatility, idiosyncratic volatility and book-to-market) and the Fama-French six equity factors. These results are presented in Table 15.

The first column of Table 15 demonstrates SSD_o provides information that is independent from a factor generated using Altman's Z-score. It is worth noting that regressing a Z-score option factor on SSD_o produces a statistically insignificant alpha, indicating SSD_o subsumes an Altman Z option factor. Risk-adjusted returns are larger than raw returns as evidenced by regressing SSD_o on a set of six factors from both Cao et al. (*forthcoming*) and Horenstein et al. (2020). Regressing SSD_o on the six Fama-French equity factors results in huge statistically significant risk-adjusted returns of 2.92% per month. Overall, Table 15 provides strong evidence that SSD_o provides information which is orthogonal to existing option and equity factors.

Figure 7: Cumulative Returns to SSD_o Investment Strategy: 1996 - 2017

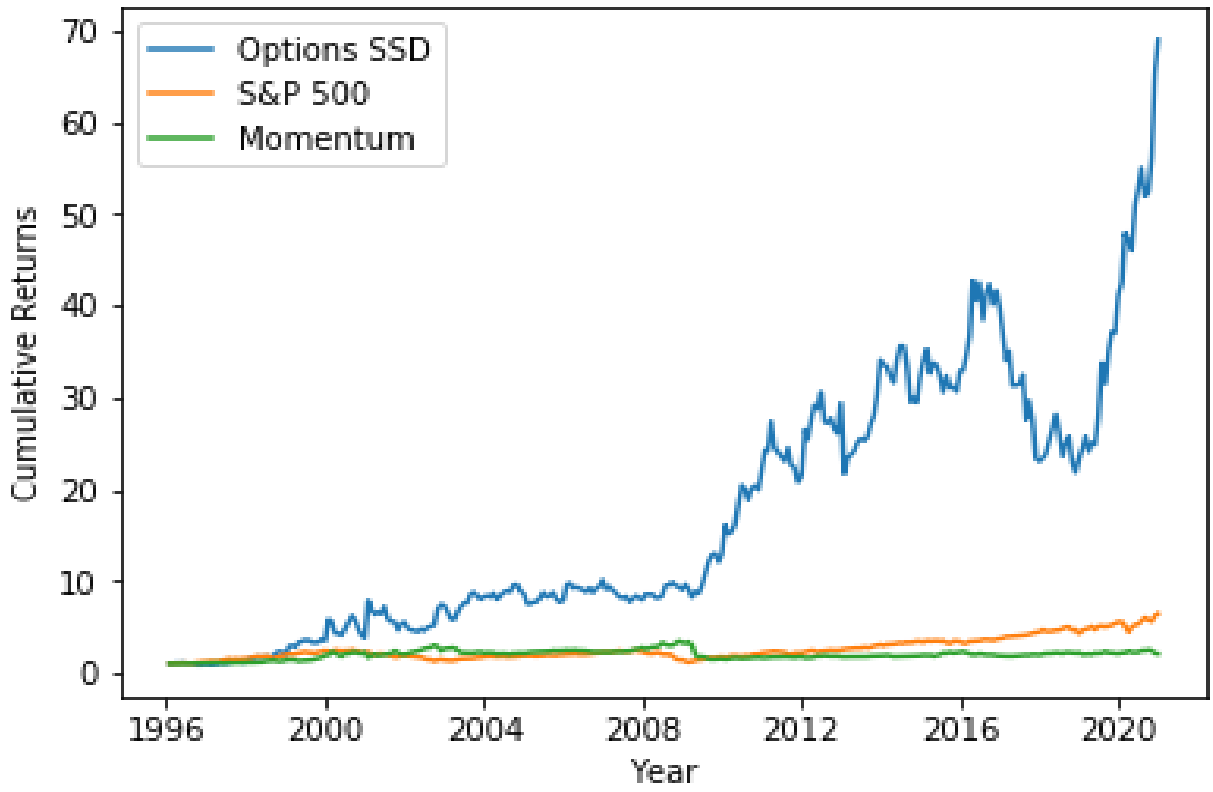


Figure 7 is the call option analogue to Figures 3 and 5 and shows the cumulative return to \$1 invested in SSD_o in January 1996 compared to \$1 invested in the S&P 500 index and \$1 invested in momentum in the same month. In Appendix E I present a graphical comparison of SSD_o with call option factors generated from both Altman's Z-Score and a WLS update

of the Z-Score coefficients. An investment of $1 in 1996 would be worth almost $70 by the end of 2020. This cumulative return dwarfs both buy and hold and momentum.

4.4 Credit Default Swap Factor

Analogous to the previous three factors, I motivate the formation of the CDS factor by first looking at average returns and bankruptcy risk for the decile cross-section of CDS portfolios sorted on the random forest probability of bankruptcy.

Before describing the decile returns presented in Table 16, it is important to emphasize the differences in what "returns" mean for stocks, bonds and call options compared to CDSs. When the asset in question is a stock, bond or call option, monthly returns represent the percentage difference between the price an investor can purchase the asset for and the price an investor can sell the asset for after a period of one month. CDS prices, on the other hand, are denominated in basis points. They are the price, in basis points, of the debt position being insured which the buyer of CDS protection pays the seller of CDS protection annually[11]. Therefore, CDS returns indicate the percent difference in the basis point rate of purchasing CDS protection in month t and then immediately selling that same level of protection in month $t+1$. Going long a CDS implies buying CDS protection, while shorting a CDS implies selling CDS protection.

Table 16: CDS Decile Mean Returns and Bankruptcy Risk

Sort	Distress (1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	Safe (10)	(1)-(10)
RF 2002	0.31%	0.38%	1.11%	2.12%	1.68%	1.52%	1.60%	1.94%	1.52%	1.58%	1.27%
t	(0.29)	(0.44)	(1.16)	(2.05)	(1.97)	(1.87)	(1.75)	(2.22)	(1.57)	(1.60)	(1.46)
$\bar{Z}_{RF}$	0.85	0.70	0.48	0.28	0.18	0.10	0.06	0.03	0.02	0.01	

Note: This table presents mean returns for various portfolios formed by sorting firms based on my random forest produced probability of bankruptcy. The row RF 1996 presents mean returns for each decile. t-statistics are in parentheses. $\bar{Z}$ represents the mean of the bankruptcy risk measure used for each portfolio.

CDS returns are much more volatile than stock, bond or call option returns, as evidenced by the small t-statistics in Table 16 — only three of the decile portfolios have statistically significant returns. Despite this, there is a much clearer pattern of increasing returns to safety than exists in the cross-section of equity returns. This reflects investor overreaction to changes in perceived risk for firms which were considered safe, while similarly sized changes in perceived risk do not impact firms which investors already deemed risky.

There are unique challenges to the formation of SSD_c that do not exist for the other three bankruptcy risk factors. While there are only approximately 10% as many firms with call option contracts compared to firms with exchange listed common stock, the cross-section of CDS contracts is formed using only the 237 firms in the S&P 500 index which have CDS contracts trade at any point between 2002 and 2020. On average there are only 76 CDS contracts trading each month (meaning decile portfolios are formed using, on average, seven CDS contracts), and some months there are as few as 3 actively traded CDS contracts. This

[11] Although payments are usually made each quarter.

makes portfolio formation using only a single decile impractical, even for firms with high levels of bankruptcy risk.

Given this portfolio size limitation and the fact that firms large enough to be in the S&P 500 index fail at a much lower rate than other firms regardless of the random forest estimated bankruptcy risk, SSD_c is formed as follows. In June of each year firms are sorted on my random forest estimated probability of bankruptcy using accounting and market equity information released in December of the previous year. SSD_c is formed by going long an equal weighted average of CDS contracts in the three safest deciles and short an equal weighted average of CDS contracts in the three riskiest deciles. The direction of this sort is consistent with both SSD_e and Friewald et al. (2014). Table 17 shows summary statistics related to SSD_c.

Since SSD_c — like SSD_o — is constructed using a univariate instead of a bivariate sort there could again be a concern that the results are driven by size instead of bankruptcy risk given prominence of size in predicting bankruptcy. Since there are no common CDS factors for me to regress SSD_c on, I again emphasize turnover in the bankruptcy risk deciles is nearly three times larger than turnover in the size deciles. This implies the results are unlikely to be driven by size.

Table 17: Bankruptcy Risk CDS Factor: 2002 - 2020

	High (1-3)	Mid (4-7)	Low (8-10)	SSD_c	t(SSD_c)	Max Drawdown	SR
SSD_c	0.83%	1.65%	1.71%	1.04%	2.55	8.9%	0.56

Note: This table presents information related to the option bankruptcy risk factor. SSD_c is formed by sorting firms in June on my random forest estimated probability of bankruptcy and going long an equal weighted average of the three safest deciles and going short an equal weighted average of the three riskiest deciles. SSD_c is the mean monthly returns from the long-short bankruptcy risk factor, t(SSD_o) is the monthly t-statistic associated with the factor. Max drawdown is the maximum peak-to-trough decrease over the full 22 year sample without regard to a specific time period. SR is the annualized Sharpe ratio. The number is parentheses ((6-8) for example) represent the component decile portfolios from Table 13.

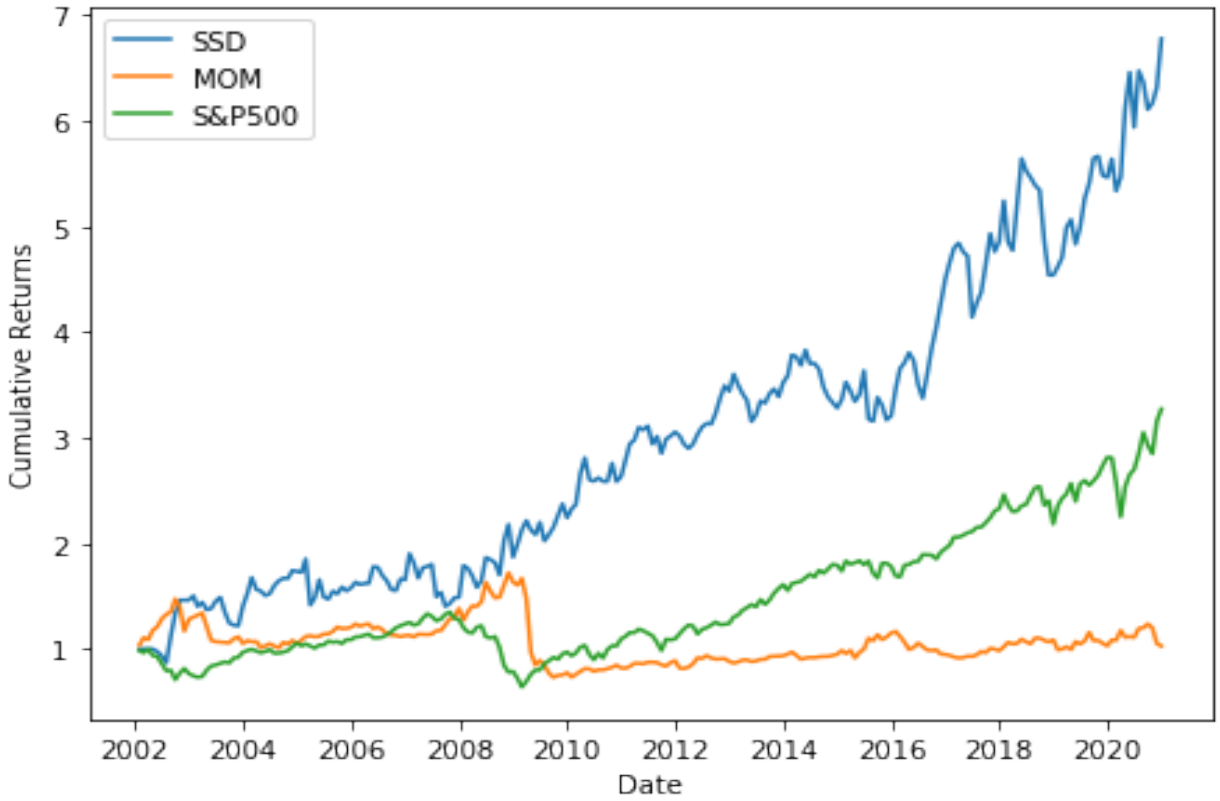
The return on SSD_c is 1.04% per month, on average and is statistically significant with a t-statistic of 2.55. The maximum drawdown is also favorable compared to both a buy and hold return (59%) and SSD_e and SSD_o (29.4% and 46.8%, respectively).

Figure 8 shows the cumulative return to $1 invested in SSD_c in October 2002 compared to the momentum strategy and a buy and hold strategy. I present an equivalent graphical comparison between SSD_c formed on my random forest estimated probability of bankruptcy and CDS factors generated using Altman's Z-score and a WLS update of the Z-score coefficients in Appendix E. Although CDS data exists for a limited number of firms beginning in 2001, there were not enough firm with debt insured by CDS contracts to form SSD_c until October 2002.

There are three notable features of the cumulative return series in Figure 8. First, SSD_c is able to distinguish between firms with low but increasing levels of bankruptcy risk (i.e.

those firms with rapidly increasing debt insurance costs), and firms with bankruptcy risk so high that CDS contracts do not get materially more expensive during periods of economic downturns. This is particularly evident during the Great Recession where returns to SSD_c continue to increase despite bankruptcy risk increasing across the entire cross-section of CDS contracts. Second, SSD_c has a convex shape and is becoming more profitable as CDS contracts become more liquid. Over the course of only 19 years 1$ invested in SSD_c increased in value almost 800%. Lastly, the fact that SSD_c beats the buy and hold return is noteworthy since they are formed using the same firms (recall SSD_c is formed using only firms from the S&P 500 index).

Figure 8: Cumulative Returns to DMS_c Investment Strategy: 2002 - 2020



4.5 Updating the Factor

Just as it is unreasonable to expect the risk factors included in previous bankruptcy risk models — including Altman's Z-score — to remain the dominant predictors of bankruptcy over time, it is unreasonable to expect my bankruptcy risk model in its current form to accurately predict bankruptcy *ad infinitum*. Insofar as bankruptcy risk is a predictor of asset returns across multiple markets, this means that it is unreasonable to expect my bankruptcy risk factor to generate statistically significant returns *ad infinitum*. It will eventually need to be retrained. How often, and at what points this retraining occurs are both important and difficult questions to answer.

To investigate this problem, I generate long-short characteristic sorted portfolios using a subset of bankruptcy risk models proposed from 1968 (Altman's Z-score) to 2000 (the F score of Piotroski (2000) and the revised Altman Z (2000)). I choose 2000 as the end date to allow a long enough out-of-sample period to determine how long, if at all, each bankruptcy risk model generates returns post-publication. I use the average duration from publication to the date the factor no longer becomes an effective predictor of returns as an indicator of how long SSD_i (for $i = e, b, o, c$) is expected to be an effective investment tool without retraining. The date at which the factor is no longer "effective" is either the date of a peak in cumulative returns, or the beginning of an extended period of near zero returns.

On average each of the constructed factors takes approximately 21 years post publication before it is no longer an effective investment vehicle. For SSD_i (for $i = e, b, o, c$) this means the patterns observed in this paper can be expected to last for more than 20 years without the underlying bankruptcy risk model being updated. Of course, given the availability of data and cheap processing power, the underlying bankruptcy risk model can be updated on a yearly basis. As I have demonstrated in this paper it appears that bankruptcy risk can predict returns in multiple markets. Being able to more accurately predict bankruptcy risk by expanding the number of bankrupt observations in the underlying model each year could increase the effectiveness of the bankruptcy risk factor. One of the benefits of the two-pass random forest framework used to generate my bankruptcy probability is that it easily allows for additional observations and additional variables to be included each time the model is updated.

5 Conclusion

Despite the well-known equilibrium theoretic result that higher average returns are compensation to investors for holding assets with higher levels of risk, empirically this is not always true. Lower risk equities and credit default swaps on average earn higher returns, while riskier bonds and options on average earn higher returns. Motivated by this empirical fact, I propose a new factor sorted on the probability of bankruptcy as a direct proxy for risk. Instead of making the subjective choice of which among the myriad bankruptcy risk models proposed by the literature to use, I introduce a new model of firm bankruptcy built as a sequence of two random forests — the first to select a parsimonious set of predictors from nearly 250 proposed bankruptcy risk indicators, and the second to classify firms.

The probability of bankruptcy generated by my random forest has predictive power in four markets — equity markets (SSD_e), bond markets (SSD_b), options markets (SSD_o) and CDS markets (SSD_c). SSD_e earns 0.23% per month and cannot be explained by the CAPM or Fama-French five-factor model. Additionally, SSD_e provides incremental explanatory power for the cross-section of returns in the presence of the factor zoo when tested using the two-step method of Feng, Giglio and Xiu (2020). SSD_b earns 0.15% per month and cannot be explained by the excess bond market return, six popular equity factors, six popular bond factors, or the union of bond and equity factors. Likewise, SSD_o — which earns 1.97% per

month — subsumes the existing Altman Z option factor and cannot be explained by six popular equity factors or six popular options factors. SSD_c also earns 1.04% per month, although there are no existing popular CDS factors to compare it to.

Just as it is unrealistic to expect the same sources of risk that most impacted firms in the 1960s to be the same sources of risk impacting firms today, it is unrealistic to think my random forest model in its current form uses risk factors that will best predict bankruptcy *ad infinitum*. Although long-short factors generated by sorting on the output of other bankruptcy risk models successfully predict returns for an average of 20 years post-publication, it may be beneficial for an investor to update the underlying bankruptcy risk model annually to maximize returns. The two step procedure used in this paper provides a framework by which the underlying bankruptcy risk model can be updated to both include new predictors proposed by the literature and add data points as more firms go bankrupt over time.

Bibliography

Abarbanell, Jeffrey; Bushee, Brian. "Abnormal Returns to a Fundamental Analysis Strategy". *The Accounting Review*. Vol 73. No 1 (1998). pp 19 - 45.

Acosta-Gonzalez, Eduardo; Fernandez-Rodriguez, Fernando. "Forecasting Financial Failure of Firms Via Genetic Algorithms". *Computational Economics*. Vol 43. Issue 2 (2014). pp 133 - 157.

Agarwal, Vineet; Taffler, Richard. "Comparing the Performance of Market-Based and Accounting-Based Bankruptcy Prediction Models". *Journal of Banking & Finance*. Vol 32. Issue 8 (2008). pp 1541 - 1551.

Aharony, Joseph; Jones, Charles; Swary, Itzhak. "An Analysis of Risk and Return Characteristics of Corporate Bankruptcy Using Capital Market Data". *The Journal of Finance*. Vol 35. Issue 4 (1980). pp 1001 - 1016.

Aktas, Nihat; de Bodt, Eric; Lobez, Frèdèric; Statnik, Jean-Christophe. "The Information Content of Trade Credit". *Journal of Banking & Finance*. Vol 36. Issue 5 (2012). pp 1402 - 1413.

Allayannis, George; Brown, Gregory; Klapper, Leora. "Capital Structure and Financial Risk: Evidence From Foreign Debt Use in East Asia". *The Journal of Finance*. Vol 58. No 6 (2003). pp 2667 - 2710.

Almaida, Caio; Freire, Gustavo. "Pricing of index options in incomplete markets". *Journal of Financial Economics*. In Press. `https://doi.org/10.1016/j.jfineco.2021.05.041` 2021.

Altman, Edward. "Financial Ratios, Discriminant Analysis and the Prediction of Corporate

Bankruptcy". *The Journal of Finance*. Vol 23. No 4 (1968). pp 589 - 609.

Altman, Edward. "Predicting Railroad Bankruptcies in America". *Bell Journal of Economics and Management Science*. Vol 4. No 1 (1973). pp 184 - 211.

Altman, Edward; Loris, Bettina. "A Financial Early Warning System for Over-the-Counter Broker-Dealers". *The Journal of Finance*. Vol 31. Issue 4 (1976). pp 1201 - 1217.

Altman, Edward. "Predicting Performance in the Savings and Loan Association Industry". *Journal of Monetary Economics*. Vol 3. Issue 4 (1977). pp 443 - 466.

Altman, Edward; Haldeman, Robert; Narayanan, Paul. "Zeta Analysis: A New Model to Identify Bankruptcy Risk of Corporations". *Journal of Banking and Finance*. Vol 1. Issue 1 (1977). pp 29 - 54.

Altman, Edward. "Why Businesses Fail". *Journal of Business Strategy*. Vol 3. Issue 4 (1983). pp 15 - 21.

Altman, Edward. "Predicting Financial Distress of Companies: Revisiting the Z-Score and Zeta® Models". *NYU Stern Working Paper*. 2000.

Altman, Edward; Iwanicz-Drozdowska, Malgorzata; Laitinen, Erkki; Suvas, Arto. "Financial Distress Prediction in an International Context: A Review and Empirical Analysis of Altman's Z-Score Model". *Journal of International Financial Management & Accounting*. Vol 28. Issue 2 (2017). pp 131 - 171.

Alwathainani, Abdulaziz. "Consistency of Firm's Past Financial Performance Measures and Future Returns". *The British Accounting Review*. Vol 41. Issue 3 (2009). pp 184 - 196.

Amihud, Yakov; Mendelson, Haim. "Asset Pricing and the Bid-Ask Spread". *Journal of Financial Economics*. Vol 17. No 2 (1986). pp 223 - 249.

Andersen, Torben; Fusari, Nicola; Todorov, Viktor; Varneskov, Rasmus. "Spatial Dependence in Option Observation Errors". *Econometric Theory*. Vol 37. Issue 2 (2021). pp 205 - 247.

Asness, Clifford; Frazzini, Andrea; Pedersen, Lasse. "Quality Minus Junk". *Review of Accounting Studies*. Vol 24 (2019). pp 34 - 112.

Asness, Clifford; Moskowitz, Tobias; Pedersen, Lasse. "Value and Momentum Everywhere". *The Journal of Finance*. Vol 68. Issue 3 (2013). pp 929 - 985.

Asquith, Paul; Covert, Thom; Pathak, Parag. "The Effects of Mandatory Transparency in Financial Market Design: Evidence from the Corporate Bond Market". *NBER Working Paper No 19417*. 2019.

Avramov, Doron; Chordia, Tarun; Jostova, Gergana; Philipov, Alexander. "Dispersion in Analysts' Earnings Forecasts and Credit Rating". *Journal of Financial Economics*. Vol 91. Issue 1 (2009). pp 83 - 101.

Bai, Jennie; Bali, Turan; Wen, Quan. "Common Risk Factors in the Cross-Section of Corporate Bond Returns". *Journal of Financial Economics*. Vol 131. Issue 3 (2019). pp 619 - 642.

Bakshi, Gurdip; Kapadia, Nikunj. "Delta-Hedged Gains and the Negative Market Volatility Risk Premium". *The Review of Financial Studies*. Vol 16. Issue 2 (2003). pp 527 - 566.

Bao, Jack; Pan, Jun; Wang, Jiang. "The Illiquidity of Corporate Bonds". *The Journal of Finance*. Vol 66. Issue 3 (2011). pp 911 - 946.

Barry, Christopher; Brown, Stephen. "Differential Information and the Small Firm Effect". *Journal of Financial Economics*. Vol 13. Issue 2 (1984). pp 283 - 294.

Basu, Shankar. "Investment Performance of Common Stock in Relation to Their Price-Earnings Ratios: A Test of the Efficient Market Hypothesis". *The Journal of Finance*. Vol 32. Issue 3 (1977). pp 663 - 682.

Beaver, William. "Financial Ratios as Predictors of Failure". *Journal of Accounting Research*. Vol 4 (1966). pp 71 - 111.

Beaver, William. "Market Prices, Financial Ratios, and the Prediction of Failure". *Journal of Accounting Research*. Vol 6. No 2 (1968). pp 179 - 192.

Bennett, Paul; Wei, Li, "Market Structure, Fragmentation, and Market Quality". *Journal of Financial Markets*. Vol 9. Issue 1 (2006). pp 49 - 78.

Breiman, Leo. "Random Forests". *Machine Learning*. Vol 45 (2001). pp 5 - 32.

Breiman, Leo; Friedman, Jerome; Olshen, Richard; Stone, Charles. "Classification and Regression Trees". Wadsworth, New York. 1984.

Brooks, Robert; Chance, Don; Shafaati, Mobina. "The Cross-Section of Individual Equity Option Returns". *LSU Working Paper*. 2018.

Campbell, John; Hilscher, Jens; Szilagyi, Jan. "In Search of Distress Risk". *The Journal of Finance*. Vol 63. Issue 6 (2008). pp 2899 - 2939.

Cao, Jie; Han, Bing; Zhan, Xintong; Tong, Qing. "Option Return Predictability". *Review of Financial Studies*. Forthcoming.

Chan, Louis; Lakonishok, Josef; Sougiannis, Theodore. "The Stock Market Valuation of Research and Development Expenditures". *The Journal of Finance*. Vol 56. Issue 6 (2001). pp 2431 - 2456.

Chen, Andrew; Zimmermann, Tim. "Open Source Cross-Sectional Asset Pricing". *Critical Finance Review. Forthcoming.*

Chen, Huafeng; Kacpercyk, Marcin; Ortiz-Molina. "Do Nonfinancial Stakeholders Affect the Pricing of Risky Debt? Evidence from Unionized Workers". *Review of Finance*. Vol 16. Issue 2 (2012). pp 347 - 383.

Chordia, Tarun; Goyal, Amit; Nozawa, Yoshio; Subrahmanyam, Avanidhar; Tong, Qing. "Are Capital Market Anomalies Common to Equity and Corporate Bond Markets?" *Journal of Financial and Quantitative Analysis*. Vol 52. No 4 (2017). pp 1301 - 1342.

Christoffersen, Peter; Fournier, Mathieu; Jacobs, Kris. "The Factor Structure in Equity Options". *The Review of Financial Studies*. Vol 31. Issue 2 (2018). pp 595 - 637.

Cici, Gjergji; Gibson, Scott; Moussawi, Rabih. "Explaining and Benchmarking Corporate Bond Returns". SSRN ID 2995626. 2017.

Clayton, Matthew; Ravid, S. Abraham. "The Effect of Leverage on Bidding Behavior: Theory and Evidence from the FCC Auctions". *The Review of Financial Studies*. Vol 15. Issue 3 (2002). pp 723 - 750.

Crawford, Steven; Perotti, Pietro; Price III, Richard; Skousen, Christopher. "Financial Statement Anomalies in the Bond Market". *Financial Analysts Journal*. Vol 75. Issue 3 (2019). pp 105 - 124.

Daniel, Kent; Titman, Sheridan. "Market Reactions to Tangible and Intangible Information". *The Journal of Finance*. Vol 61. Issue 4 (2006). pp 1605 - 1643.

DeAngelo, Harry; DeAngelo, Linda. "Dividend Policy and Financial Distress: An Empirical Investigation of Troubled NYSE Firms". *The Journal of Finance*. Vol 45. Issue 5 (1990). pp 1415 - 1431.

Dichev, Ilia. "Is the Risk of Bankruptcy a Systemic Risk?" *The Journal of Finance*. Vol 53. Issue 3 (1998). pp 1131 - 1147.

Dick-Nielsen, Jens. "How to Clean Enhanced TRACE Data". SSRN ID 2337908. 2014.

Dick-Nielsen, Jens. "Liquidity Biases in TRACE". *The Journal of Fixed Income*. Vol 19. No 2 (2009). pp 34 - 55.

Edminster, Robert. "An Empirical Test of Financial Ratio Analysis for Small Business

Failure Prediction". *Journal of Financial and Quantitative Analysis*. Vol 7. Issue 2 (1972). pp 1477 - 1493.

Efron, Bradley; Hastie, Trevor; Johnstone, Iain; Tibshirani, Robert. "Least Angle Regression". *The Annals of Statistics*. Vol 32. No 2 (2004). pp 407 - 499.

Elton, Edwin; Gruber, Martin; Blake, Christopher. "Fundamental Economic Variables, Expected Returns, and Bond Fund Performance". *The Journal of Finance*. Vol 50. Issue 4 (1995). pp 1229 - 1256.

Fama, Eugene; French, Kenneth. "A Five Factor Asset Pricing Model". *Journal of Financial Economics*. Vol 116. Issue 1 (2015). pp 1 - 22.

Fama, Eugene; French, Kenneth. "Common Risk Factors in the Returns on Stocks and Bonds". *Journal of Financial Economics*. Vol 33. Issue 1 (1993). pp 3 - 56.

Fama, Eugene; French, Kenneth. "The Cross-Section of Expected Stock Returns". *The Journal of Finance*. Vol 47. No 2 (1992). pp 427 - 465.

Fama, Eugene; MacBeth, James. "Risk, Return, and Equilibrium: Empirical Tests". *Journal of Politicl Economy*. Vol 81. No 3 (1973). pp 607 - 636.

Feng, Guanhao; Giglio, Stefano; Xiu, Dacheng. "Taming the Factor Zoo: A Test of New Factors". *The Journal of Finance*. Vol 75. Issue 3 (2020). pp 1327 - 1370.

Fernàndez, Albert; Garcia, Salvador; Galar, Mikel; Prati, Ronaldo; Krawczyk, Bartosz; Herrera, Francisco. "Learning From Imbalanced Data Sets". First Edition. *Springer*. Berlin, Germany. 2018.

Fons, Jerome. "The Default Premium and Corporate Bond Experience". *The Journal of Finance*. Vol 42. Issue 1 (1987). pp 81 - 97.

Foster, George; Olsen, Chris; Shevlin, Terry. "Earnings Releases, Anomalies, and the Behavior of Security Returns". *The Accounting Review*. Vol 59. No 4 (1984). pp 574 - 603.

Franzen, Laurel; Rodgers, Kimberly; Simin; Timothy. "Measuring Distress Risk: The Effect of R&D Intensity". *The Journal of Finance*. Vol 62. Issue 6 (2007). pp 2931 - 2967.

Frazzini, Andrea; Pedersen, Lasse. "Betting Against Beta". *Journal of Financial Economics*. Vol 111. Issue 1 (2014). pp 1 - 25.

Friewald, Nils; Wagner, Christian; Zechner, Josef. "The Cross-Section of Credit Risk Premia and Equity Returns". *The Journal of Finance*. Vol 69. Issue 6 (2014). pp 2419 - 2469.

Gentry, James; Newbold, Paul; Whitford, David. "Classifying Bankrupt Firms with Funds

Flow Components". *Journal of Accounting Research*. Vol 23. No 1 (1985). pp 146 - 160.

Gharghori, Philip; Maberly, Edwin; Nguyen, Annette. "Informed Trading Around Stock Split Announcements: Evidence From the Options Market". *Journal of Financial and Quantitative Analysis*. Vol 52. Issue 2 (2017). pp 705 - 735.

Goyal, Amit; Saretto, Alessio. "Cross-Section of Options Returns and Volatility". *Journal of Financial Economics*. Vol 94. Issue 2 (2009). pp 310 - 326.

Griffin, John; Lemmon, Michael. "Book-to-Market Equity, Distress Risk, and Stock Returns". *The Journal of Finance*. Vol 57. Issue 5 (2002). pp 2317 - 2336.

Gu, Shihao; Kelly, Bryan; Xiu, Dacheng. "Empirical Asset Pricing Via Machine Learning". *Review of Financial Studies*. Vol 33. Issue 5 (2020). pp 2223 - 2272.

Hartzmark, Samuel; Salomon, David. "The Dividend Month Premium". *Journal of Financial Economics*. Vol 109. Issue 3 (2013). pp 640 - 660.

Harvey, Campbell; Liu, Yan. "A Census of the Factor Zoo". SSRN ID 3341728. 2019.

Haugen, Robert; Baker, Nardin. "Commonality in the Determinants of Expected Stock Returns". *Journal of Financial Economics*. Vol 41. Issue 3 (1996). pp 401 - 439.

Hillegeist, Stephen; Keating, Elizabeth; Cram, Donald; Lundstedt, Kyle. "Assessing the Probability of Bankruptcy". *Review of Accounting Studies*. Vol 9 (2004). pp 5 - 34.

Hirschleifer, David; Hsu, Po-Hsuan; Li, Dongmei. "Innovative Efficiency and Stock Returns". *Journal of Financial Economics*. Vol 107. Issue 3 (2013). pp 632 - 654.

Ho, Chun-Yu; McCarthy, Patrick; Yang, Yi; Ye, Xuan. "Bankruptcy in the Pulp and Paper Industry: Market's Reaction and Prediction". *Empirical Economics*. Vol 45 (2013). pp 1205 - 1232.

Hopwood, William; McKeown, James; Mutchler, Jane. "A Reexamination of Auditor Versus Model Accuracy Within the Context of the Going-Concern Opinion Decisions". *Contemporary Accounting Research*. Vol 10. Issue 2 (1994). pp 409 - 431.

Horenstein, Alex; Vasquez, Aurelio; Xiao, Xiao. "Common Factors in Equity Option Returns". *SSRN ID 3290363*. 2020.

Hou, Kewei. "Industry Information Diffusion and the Lead-Lag Effect in Stock Returns". *The Review of Financial Studies*. Vol 20. Issue 4 (2007). pp 1113 - 1138.

Hou, Kewei; Xue, Chen; Zhang, Lu. "Digesting Anomalies: An Investment Approach". *Review of Financial Studies*. Vol 28. Issue 3 (2015). pp 650 - 705.

Hou, Kewei; Robinson, David. "Industry Concentration and Average Stock Returns". *The Journal of Finance*. Vol 61. Issue 4 (2006). pp 1927 - 1956.

Huang, Alan. "The Cross-Section of Cash Flow Volatility and Expected Stock Returns". *Journal of Empirical Finance*. Vol 16. Issue 3 (2009). pp 409 - 429.

Jensen, Theis; Kelly, Bryan; Pedersen, Lasse. "Is There a Replication Crisis in Finance?" *NBER Working Paper No 28432*. 2021.

Jones, Stewart; Hensher, David. "Predicting Firm Financial Distress: A Mixed Logit Model". *The Accounting Review*. Vol 79. No 4 (2004). pp 1011 - 1038.

Jostova, Gergana; Nikolova, Stanislava; Philipov, Alexander; Stahel, Christof. "Momentum in Corporate Bond Returns". *The Review of Financial Studies*. Vol 26. Issue 7 (2013). pp 1649 - 1693.

Kieschnick, Robert; Laplante, Mark; Moussawi, Rabih. "Working Capital Management and Shareholders' Wealth". *Review of Finance*. Vol 17. Issue 5 (2013). pp 1827 - 1852.

Kimball, Miles. "Precautionary Saving in the Small and in the Large". *Econometrica*. Vol 58 (1990). pp 53 - 73.

King, Gary; Zeng, Langche. "Logistic Regression in Rare Events Data". *Political Analysis*. Vol 9. Issue 2 (2001). pp 137 - 163.

Korobow, Leon; Stuhr, David. "Toward Early Warning of Changes in Banks' Financial Condition: A Progress Report". *Federal Reserve Bank of New York Monthly Review* (1975). pp 157 - 165.

Lee, Shih-Cheng; Chen, Jiun-Lin; Jiang, I-Ming; Hsu, Cheng-Yi. "Accounting Conservatism and Bankruptcy". *Journal of Accounting, Finance & Management Strategy*. Vol 7. Issue 2 (2012). pp 53 - 69.

Lennox, Clive. "Identifying Failing Companies: A Re-evaluation of the Logit, Profit and DA Approaches". *Journal of Economics and Business*. Vol 51. Issue 4 (1999). pp 347 - 364.

Marais, D. A. J. "A Method of Quantifying Companies Relative Financial Strength". *Bank of England Discussion Paper No. 4*. 1979.

Martin, Daniel. "Early Warning of Bank Failure: A Logit Regression Approach". *Journal of Banking & Finance*. Vol 1. Issue 3 (1977). pp 249 - 276.

Meine, Christian; Supper, Hendrik; Weiss, Gregory. "Is Tail Risk Priced in Credit Default Swap Premia?" *Review of Finance*. Vol 20. Issue 1 (2016). pp 287 - 336.

Merton, Robert. "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates". *The Journal of Finance*. Vol 29. Issue 2 (1974). pp 449 - 470.

Moskowitz, Tobias; Ooi, Yao; Pedersen, Lasse. "Time Series Momentum". *Journal of Financial Economics*. Vol 104. Issue 2 (2012). pp 228 - 250.

Muravyev, Dmitriy; Pearson, Neil. "Options Trading Costs Are Lower Than You Think". *Review of Financial Studies*. Vol 33. Issue 11 (2020). pp 4973 - 5014.

Novy-Marx, Robert. "The Other Side of Value: The Gross Profitability Premium". *Journal of Financial Economics*. Vol 108. Issue 1 (2013). pp 1 - 28.

Ogachi, Daniel; Ndege, Richard; Gaturu, Peter; Zoltan, Zeman. "Corporate Bankruptcy Prediction Model, a Special Focus on Listed Companies in Kenya". *Journal of Risk and Financial Management*. Vol 13. Issue 3 (2020).

Ohlson, James. "Financial Ratios and the Probabilistic Prediction of Bankruptcy". *Journal of Accounting Research*. Vol 18. No 1 (1980). pp 109 - 131.

Penman, Stephen; Richardson, Scott; Tuna, Irem. "The Book-to-Price Effect in Stock Returns: Accounting for Leverage". *Journal of Accounting Research*. Vol 45. Issue 2 (2007). pp 427 - 467.

Philosophov, Leonid; Batten, Jonathan; Philosophov, Vladimir. "Predicting the Event and Time Horizon of Bankruptcy Using Financial Ratios and the Maturity Schedule of Long-Term Debt". *Mathematics and Financial Economics*. Vol 1 (2008). pp 181 - 212.

Pindado, Julio; Rodrigues, Lius; de la Torre, Chabela. "Estimating Financial Distress Likelihood". *Journal of Business Research*. Vol 61. Issue 9 (2008). pp 995 - 1003.

Piotroski, Joseph. "Value Investing: The Use of Historical Financial Statement Information to Separate Winners from Losers". *Journal of Accounting Research*. Vol 38 (2000). pp 1 - 41.

Probst, Phillip; Wright, Marvin; Boulesteix, Anne-Laure. "Hyperparameters and Tuning Strategies for Random Forest". *Data Mining and Knowledge Discovery*. Vol 9. Issue 3 (2019). pp 1301 - 1316.

Rose-Green, Ena; Lovata, Linda. "The Relationship Between Firms' Characteristics in the Periods Prior to Bankruptcy Filing and Bankruptcy Outcome". *Accounting and Finance Research*. Vol 2. No 1 (2013). pp 97 - 109.

Santomero, Anthony; Vinso, Joseph. "Estimating the Probability of Failure for Commercial Banks and the Banking System". *Journal of Banking & Finance*. Vol 1. Issue 2 (1977). pp 185 - 205.

Shaked, Israel; Orelowitz, Brad. "Understanding Retail Bankruptcy". *American Bankruptcy Institute Journal*. Vol 36. Issue 11 (2017). pp 20-73.

Shumway, Tyler. "Forecasting Bankruptcy More Accurately: A Simple Hazard Model". *The Journal of Business*. Vol 74. No 1 (2001). pp 101 - 124.

Singhal, Rajeev; Zhu, Yun. "Bankruptcy Risk, Costs and Corporate Diversification". *Journal of Banking & Finance*. Vol 37. Issue 5 (2013). pp 1475 - 1489.

Sinkey Jr. Joseph. "A Multivariate Statistical Analysis of the Characteristics of Problem Banks". *The Journal of Finance*. Vol 30. Issue 1 (1975). pp 21 - 36.

Soliman, Mark. "The Use of DuPont Analysis by Market Participants". *The Accounting Review*. Vol 83. No 3 (2008). pp 823 - 853.

Sun, Lili. "A Re-Evaluation of Auditors' Opinions Versus Statistical Models in Bankruptcy Prediction". *Review of Quantitative Finance and Accounting*. Vol 27 (2007). pp 55 - 78.

Sweeney, Richard; Warga, Arthur. "The Pricing of Interest-Rate Risk: Evidence from the Stock Market". *Journal of Finance*. Vol 41. No 2 (1986). pp 393 - 410.

Taffler, Richard. "Going, Going, Gone — Four Factors that Predict". *Accountancy*. Vol 88 (1977). pp 50 - 54.

Tamari, Meir. "Financial Ratios as a Means of Predicting Bankruptcy". *Management International Review*. Vol 6. No 4 (1966). pp 15 - 21.

Thomas, Jacob; Zhang, Frank. "Tax Expense Momentum". *Journal of Accounting Research*. Vol 49. Issue 3 (2011). pp 791 - 821.

Tibshirani, Robert; "Regression Shrinkage and Selection Via the Lasso". *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*. vol 58. Issue 1 (1996). pp 267 - 288.

Todorov, Viktor. "Nonparametric Spot Volatility from Options". *The Annals of Applied Probability*. Vol 29. No 6 (2019). pp 3590 - 3636.

Warner, Jerold. "Bankruptcy Costs: Some Evidence". *The Journal of Finance*. Vol 32. No 2 (1977). pp 337 - 347.

Wilcox, Jarrod. "A Simple Theory of Financial Ratios as Predictors of Bankruptcy". *Journal of Accounting Research*. Vol 9. No 2 (1971). pp 389 - 395.

Wilcox, Jarrod. "The Gamblers Ruin Approach to Business Risk". *Sloan Management*

Review. Vol 18. Issue 1 (1976). pp 33 - 46.

Vassalou, Maria; Xing, Yuhang. "Default Risk in Equity Returns". *The Journal of Finance*. Vol 59. No 2 (2004). pp 831 - 868.

Zmijewski, Mark. "Methodological Issues Related to the Estimation of Financial Distress Prediction Models". *Journal of Accounting Research*. Vol 22 (1984). pp 59 - 82.

Appendix

Appendix A - Random Forest

The random forest was originally introduced by Breiman (2001). It is an iterative collection of classification and regression tree (CART) algorithms, which were themselves originally developed for machine learning purposes by Brieman et al. (1984). Because random forests are made up of a series of classification and regression trees, they belong to a group of models known as ensemble models. The predictions of individual trees, known as "base learners" are aggregated to make the final prediction. If the random forest is used for classification, predictions are made according to majority rule. If the random forest is used for regression, predictions are averaged over the component trees. Because the prediction of so many base learners are combined, predictions made by random forests are much more stable than predictions of individual trees. In this way the random forest model trades off bias in the prediction for a drastic reduction in the variance.

Each tree in the forest is grown according to the CART algorithm, which can be summarized as follows. Beginning at the root node, the data is split (splits are always binary) on a variable chosen from a bootstrap random sample of the full list of explanatory variables. The resulting two groups are called "branches". The split criteria is selected as to make each resulting group (branch) as homogeneous as possible, where homogeneity is defined with respect to an "impurity function". The impurity function is essentially the loss function to be minimized. The process of making these splits is known as "growing the tree", and splits are not required to be unique. For example, the first split and the third split could be on the same variable if further splitting on an already used variable minimizes the impurity function.

The CART algorithm continues to split the data until a stopping criteria is reached. Examples of stopping criteria include a maximum number of layers (steps), a minimum number of observations in each group after a split and a minimum value of the impurity function. Splits are made in a greedy manner. Greedy algorithms choose paths which are locally optimal, and are therefore not guaranteed to arrive at a globally optimal solution. In practice trees are grown such that they are overfit and then "pruned". This process is undertaken because it is possible for a branch close to the root node to not reduce the impurity function by a meaningful amount, but have a child node that splits the data in a way that significantly improves the tree's predictive accuracy. Pruning the tree in this way ensures these child nodes are reached. The terminal nodes of the tree are known as "leaf" nodes. Predictions

are made according to the mean value in each leaf node.

Formally, random forests can be represented mathematically using the following indicator function

$$f(x_i, \phi, N, K, L) = \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} \phi_k \mathbf{1}_{x_i \in C_k(L)} \tag{3}$$

where x_i is the raw input, ϕ is the mean value of the dependent variable in leaf node k, N is the number of trees in the forest, and K and L describe each tree's number of leaf nodes and depth. The indicator function makes clear that each data point input into the model can only reach one leaf node (leaf node k).

Impurity functions can be any function. However, common impurity functions include the regular L_2 loss function

$$L(\phi, C) = \frac{1}{|C|} \sum_{x_i \in C} (x_i - \phi)^2 \tag{4}$$

where C is the number of observations which fall in leaf node k and the remaining variables and parameters are defined as before.

Appendix B - Lasso

The least absolute shrinkage and selection operator (lasso) was developed by Tibshirani in 1996. It is a regularized regression which shrinks OLS coefficients towards zero and allows a subset of those coefficients to equal zero — thus acting as a variable selection tool.

Formally, suppose there is a set of data with T observations and P predictors. Further suppose that $\boldsymbol{y} = (y_1, ..., y_T)^T$ are the dependent variables (called responses in the machine learning literature) and $\boldsymbol{X} = (\boldsymbol{x_1}|...|\boldsymbol{x_P})$ is the matrix of independent variables (called predictors in the machine learning literature) where each $\boldsymbol{x_j}$ has T observations. The lasso solves the objective function

$$(\hat{\boldsymbol{\beta}}, \lambda) = \arg\min ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1 \tag{5}$$

where λ is the regularization term penalizing model complexity, $||\boldsymbol{\beta}||_1 = \sum_{j=1}^{P} |\beta|$ is the L_1 norm and $||\boldsymbol{\beta}||_2 = \sqrt{\sum_{j=1}^{P} \beta^2}$ is the L_2 norm. It is clear from equation 5 that higher levels of λ are associated with stronger regularization. When $\lambda = \infty$ all coefficients are regularized to zero and the mean value of y is the predicted value for all inputs. When $\lambda = 0$ the lasso reverts to ordinary least squares.

There is no closed form solution to the lasso objective function, and the numerical solution — efficiently obtained using the LARS algorithm (Efron et al. (2004)) — depends on the strength of the regularization. The strength of the regularization is chosen in a data driven way using cross-validation where the lasso is prevented from overfitting by testing the pseudo out-of-sample predictive ability on data not used to train the model. This pseudo out-of-sample data is rotated at each iteration of the cross-validation process so the lasso is both trained and tested on the entire data set. The value of λ which best predicts pseudo out-of-sample, on average across all iterations, is selected as optimal.

Appendix C - Bankruptcy Predictor Variable Construction

This appendix lists the full set of bankruptcy predictors used in this study along with the original source of the variable and a brief description of its construction. Studies that emphasize the importance of components of constructed variables (i.e. current assets as a component of the current ratio), those components are also included even if they are not the primary focus of the author's paper.

Table 18: Bankruptcy Predictor Variables

Name	Source	Description
LiqAssets	Altman (1968)	Working capital scaled by total assets
CumProfit	Altman (1968)	Retained earnings scaled by total assets
ScaleEBIT	Altman (1968)	Earnings before interest and taxes scaled by total assets
Solvency	Altman (1968)	Market value of equity scaled by book value of debt
CapTurn	Altman (1968)	Sales scaled by total assets
CurrentRat	Tamari (1966)	Current assets scaled by current liabilities
QuickRatio	Tamari (1966)	Liquid current assets scaled by current liabilities
DebtEquity	Ogachi et al., (2020)	Total liabilities scaled by total shareholder equity
PSolvency	Altman (1983)	Book value of equity scaled by book value of total liabilities
ROA	Piotroski (2000)	Net income before extraordinary items scaled by total assets
OpCash	Piotroski (2000)	Cash flows from operation scaled by total assets
ChgROA	Piotroski (2000)	Year over year change in net income before extraordinary items scaled by total assets
AccrualRat	Piotroski (2000)	Net income in excess of cash flows from operations scaled by total assets
ChgLev	Piotroski (2000)	Year over year change in total debt scaled by total assets
ChgCurRat	Piotroski (2000)	Year over year change in current assets scaled by current liabilities
OfferEq	Piotroski (2000)	Year over year change in shares outstanding
MargRat	Piotroski (2000)	Gross margin scaled by total assets
ChgMargRat	Piotroski (2000)	Year over year change in gross margin scaled by total assets
ChgCapTurn	Piotroski (2000)	Year over year change in sales scaled by total assets
SIZE	Ohlson (1980)	Total assets scaled by the GNP price deflator
Leverage	Ohlson (1980)	Total debt scaled by total assets
TLTA	Ohlson (1980)	Total liabilities scaled by total assets

Continued on next page

Name	Source	Description
OENEG	Ohlson (1980)	Indicator variable equal to one if total liabilities is larger than total assets
NITA	Ohlson (1980)	Net income scaled by total assets
FUTL	Ohlson (1980)	Funds from operations scaled by total liabilities
INTWO	Ohlson (1980)	Year over year percentage change in income
BEME	Fama and French (1993)	Book equity scaled by market equity
Debt	Clayton and Ravid (2002)	Total debt
IntCovRat	Clayton and Ravid (2002)	Earnings before interest and taxes scaled by total interest payments
LogSale	Clayton and Ravid (2002)	Log of total sales
Tax	Allayannis et al., (2003)	Total taxes paid
ME	Allayannis et al., (2003)	Total market equity
TanAssets	Allayannis et al., (2003)	Percentage of total assets made up of tangible assets
CFFO	Gentry et al., (1985)	Cash flows from operations
OpLoss	Hopwood et al., (1994)	Total operating expenses minus gross profits
Industry	Sun (2007)	1-digit industry SIC code
CASALES	Sun (2007)	Current assets scaled by sales
CATA	Sun (2007)	Current assets scaled by total assets
PBTCL	Taffler (1977)	Profits before taxes scaled by current liabilities
CATL	Taffler (1977)	Current assets scaled by total liabilities
CLTA	Taffler (1977)	Current liabilities scaled by total assets
NoCredInt	Taffler (1977)	Ratio of quick assets in excess of total liabilities and sales in excess of profits and depreciation, divided by 365
OCLCL	Marais (1979)	Operating cash flows scaled by current liabilities
OCLCL2	Beaver (1966)	Operating cash flows scaled by sales
CLCR	Beaver (1966)	Operating cash flows in excess of dividends paid scaled by current liabilities
LTDCR	Beaver (1966)	Operating cash flows in excess of dividends paid scaled by long term debt
DebtEq2	Warner (1977)	Total debt scaled by total shareholder equity
STDebtEq	Altman et al., (2017)	Total short-term debt scaled by total shareholder equity

Continued on next page

Name	Source	Description
IntCovRat2	Rose-Green and Lovata (2013)	The sum of operating cash flows, interest payments and taxes scaled by interest payments
DebtEbitda	Shaked and Orelowitz (2017)	Total debt scaled by earnings before interest, taxes, depreciation and amortization
GPTA	Philosophov et al., (2008)	Gross profits scaled by total assets
WorkCap	Philosophov et al., (2008)	Current assets minus current liabilities
LTDTA	Philosophov et al., (2008)	Total long-term debt scaled by total assets
PartCLRat	Philosophov et al., (2008)	Current liabilities minus long-term debt due in one year scaled by total assets
IntRat	Philosophov et al., (2008)	Interest payments scaled by total assets
MLDTA	Philosophov et al., (2008)	Long-term debt due in one year scaled by total assets
MLDTA2	Philosophov et al., (2008)	Long-term debt due in two years scaled by total assets
MLDTA3	Philosophov et al., (2008)	Long-term debt due in three years scaled by total assets
MLDTA4	Philosophov et al., (2008)	Long-term debt due in four years scaled by total assets
MLDTA5	Philosophov et al., (2008)	Long-term debt due in five years scaled by total assets
FinLoss	Pindado et al., (2008)	Earnings before interest, taxes, depreciation and amortization minus financial expenses
ChgME	Pindado et al., (2008)	Year over year change in the market value of equity
Return	Aharony et al., (1980)	One year lagged mean returns
VarRet	Aharony et al., (1980)	Year over year change in the volatility of returns
RelativeME	Shumway (2001)	Market equity scaled by the sum of all stock market equity
TradeCred	Aktas et al., (2012)	Level of trade credit as proxied by accounts payable
TobinQ	Chen et al., (2012)	Tobin's Q
Payout	Chen et al., (2012)	The sum of dividends and repurchases scaled by total assets
CFDebtRat	Beaver (1966)	Cash flows scaled by total debt

Continued on next page

Name	Source	Description
Segments	Singhal and Zhu (2013)	The number of business segments
LogAsset	Singhal and Zhu (2013)	Log of total assets
IATS	Singhal and Zhu (2013)	Intangible assets scaled by sales
NetIncome	Singhal and Zhu (2013)	Net income
LabProd	Ho et al., (2013)	Total employment scaled by total assets
WCMan	Kieschnick et al., (2013)	The sum of accounts receivable, accounts payable and inventory
ChgSale	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Year over year percentage change in sales
FinProf	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Gross profits scaled by shareholder equity
WorkCap2	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	The sum of shareholder equity, provision for risk and expenses and long-term debt minus fixed assets and other non-current assets
WorkCapReq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	The sum of subscribed shares not paid in, accrued expenses and bank loans minus the sum of short-term financial investments, cash, accrued income and current liabilities
Treasury	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Stock holdings plus cash minus loans
Equilibrium	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	The sum of shareholder equity, non-current liabilities and long-term debt scaled by total fixed assets
WCR2	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Constructed ratio "WorkCapReq" scaled by sales
Treasury2	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Constructed ratio "Treasury" scaled by sales
DebtSale	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Total debt scaled by sales

Continued on next page

Name	Source	Description
Debtness	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Liabilities scaled by the sum of liabilities and shareholder equity
EqCapRat	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Shareholder equity scaled by non-current liabilities
PayCap	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	The sum of long-term debt and current liabilities scaled by the sum of sales, depreciation and operating and investing provisions
ShareFund	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Profits scaled by shareholder equity
RetCapEmp	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Profits before interest payments scaled by the sum of shareholder equity and non-current liabilities
Margin	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Profits scaled by asset turnover
NetTurn	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Operating revenue scaled by the sum of shareholder equity and non-current liabilities
IntCover	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Operating profits scaled by interest payments
StockTurn	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Operating revenue scaled by asset turnover
ShareLiq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Shareholder equity scaled by non-current liabilities
Gearing	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	The sum of non-current liabilities and loans scaled by shareholder equity
ChgSaleSq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Squared year over year percentage change in sales
CapTurnSq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Squared capital turnover ratio

Continued on next page

Name	Source	Description
GPTASq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	The square of profits scaled by total assets
FinProfSq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	The square of profits scaled by shareholder equity
WorkCap2Sq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Constructed ratio "WorkCap2" squared
WCRSq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Constructed ratio "WorkCapReq" squared
TreasurySq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Constructed ratio "Treasury" squared
EquilSq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Constructed ratio "Equilibrium" squared
WCSaleSq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	The square of working capital scaled by sales
WCR2Sq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Constructed ratio "WCR2" squared
Treas2Sq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Constructed ratio "Treasury2" squared
DebtSaleSq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	The square of debt scaled by sales
DebtnessSq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Constructed ratio "Debtness" squared
EqCapRatSq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Constructed ratio "EqCapRat" squared
PayCapSq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Constructed ratio "PayCap" squared

Continued on next page

Name	Source	Description
QRSq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	The squared quick ratio
CRSq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	The squared current ratio
ShareFundSq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Constructed ratio "ShareFund" squared
RCESq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Constructed ratio "RetCapEmp" squared
GPTASq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	The square of profits scaled by total assets
MarginSq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	The square of profits scaled by asset turnover
NetTurnSq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Constructed ratio "NetTurn" squared
IntCoverSq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Constructed ratio "IntCover" squared
StTurnSq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Constructed ratio "StockTurn" squared
ShareLiqSq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Constructed ratio "ShareLiq" squared
GearSq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	Constructed ratio "Gearing" squared
WCTASq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	The square of working capital scaled by total assets
EBITTASq	Acosta-Gonzalez and Fernandez-Rodriquez (2014)	The square of earnings before interest and taxes scaled by total assets
NATCTC	Altman et al., (1977)	Net available for total capital scaled by total capital

Continued on next page

Name	Source	Description
SaleTC	Altman et al., (1977)	Sales scaled by total capital
EBITSale	Altman et al., (1977)	Earnings before interest and taxes scaled by sales
NATCSale	Altman et al., (1977)	Net available for total capital scaled by sales
LogTanAss	Altman et al., (1977)	Log of tangible assets
LogIntCov	Altman et al., (1977)	The log of earnings before interest and taxes scaled by interest payments
LogWCLTD	Altman et al., (1977)	Log of working capital scaled by long-term debt
WCLTD	Altman et al., (1977)	Working capital scaled by long-term debt
FCCR	Altman et al., (1977)	The sum of earnings before interest and taxes and fixed charges scaled by the sum of fixed charges before taxes and interest payments
EBITDebt	Altman et al., (1977)	Earnings before interest and taxes scaled by debt
CFFC	Altman et al., (1977)	Cash flows scaled by fixed charges
WCCE	Altman et al., (1977)	Working capital scaled by cash expenses
BETC	Altman et al., (1977)	Book equity scaled by total capital
METC	Altman et al., (1977)	Market value of equity scaled by total capital
EBITDrop	Altman et al., (1977)	Year over year change in earnings before interest and taxes
MDrop	Altman et al., (1977)	Year over year change in profits
SaleFA	Altman et al., (1977)	Sales scaled by fixed assets
DivPayRat	Wilcox (1971)	Total dividends paid scaled by net income
LiqValueIF	Wilcox (1976)	Net income minus total dividends paid
LiqValue	Wilcox (1976)	Sum of real estate, equipment and inventory
ChgAsset	Wilcox (1976)	Year over year change in total assets
QATA	Beaver (1968)	Quick assets scaled by total assets
CashTA	Beaver (1968)	Total cash holdings scaled by total assets
QACL	Beaver (1968)	Quick assets scaled by current liabilities
QASale	Beaver (1968)	Quick assets scaled by sales

Continued on next page

Name	Source	Description
WCSale	Beaver (1968)	Working capital scaled by sales
CashSale	Beaver (1968)	Total cash holdings scaled by sales
EATTA	Beaver (1966)	Earnings after taxes scaled by total assets
QuickInv	Beaver (1966)	Quick assets scaled by inventory
OpMargin	Edminster (1972)	Operating earnings scaled by revenue
InvWC	Edminster (1972)	Inventory scaled by working capital
CADebt	Edminster (1972)	Current assets scaled by total debt
FAEquity	Edminster (1972)	Fixed assets scaled by total shareholder equity
CLEquity	Edminster (1972)	Current liabilities scaled by total shareholder equity
OwnAssets	Edminster (1972)	The sum of shareholder equity and long-term debt scaled by fixed assets
InvSale	Edminster (1972)	Inventory scaled by total sales
FASale	Edminster (1972)	Fixed assets scaled by total sales
TASale	Edminster (1972)	Total assets scaled by sales
SESale	Edminster (1972)	Total shareholder equity scaled by sales
EBITSE	Edminster (1972)	Earnings before interest and taxes scaled by total shareholder equity
EBITDD	Edminster (1972)	The sum of earnings before interest and taxes and depreciation scaled by total debt
CATOR	Altman (1973)	Current assets scaled by total operating revenue
IBITTA	Altman (1973)	Income before interest and taxes scaled by total assets
RevProp	Altman (1973)	Operating revenue scaled by the value of total property
OpEff	Altman (1973)	Operating expenses scaled by operating revenue
GrowRate3	Altman (1973)	Three year percentage growth in operating revenue
LiqAss	Sinkey (1975)	The sum of cash and treasury securities scaled by total assets
LoanTA	Sinkey (1975)	Total loans scaled by total assets
OEOI	Sinkey (1975)	Operating expenses scaled by operating income
LoanRev	Sinkey (1975)	Total loans scaled by total revenue
TresRev	Sinkey (1975)	Total holdings of U.S. treasury securities scaled by revenue
IntRev	Sinkey (1975)	Total interest paid scaled by revenue
NIBTTC	Korobow and Stuhr (1975)	Net income before taxes scaled by total capital
DivTC	Korobow and Stuhr (1975)	Dividends paid scaled by total capital
BorrowTC	Korobow and Stuhr (1975)	Total borrowing scaled by total capital

Continued on next page

Name	Source	Description
TCTA	Korobow and Stuhr (1975)	Total capital scaled by total assets
OccExpTA	Korobow and Stuhr (1975)	Net occupancy expenses scaled by total assets
TresTA	Korobow and Stuhr (1975)	Total holdings of U.S. treasury securities scaled by total assets
LoanCA	Korobow and Stuhr (1975)	Loans scaled by current assets
LLTC	Martin (1977)	The sum of loans and leases scaled by total capital
CATC	Martin (1977)	Current assets scaled by total capital
NITACash	Martin (1977)	Net income scaled by the difference between total assets and cash holdings
LoanTAC	Martin (1977)	Total loans scaled by the difference between total assets and cash holdings
NLATACash	Martin (1977)	Net liquid assets scaled by the difference between total assets and cash holdings
Loans	Martin (1977)	Total loans held
NIMEA	Martin (1977)	Net interest margin scaled by earning assets
ProdEff	Martin (1977)	Non-interest expenses scaled by the difference between operating revenue and interest expenses
DivType	Martin (1977)	Common stock dividends scaled by the difference between net income and preferred stock dividends
NIATTA	Altman and Lorris (1976)	Net income after taxes scaled by total assets
LLSE	Altman and Lorris (1976)	The sum of total liabilities and loans scaled by total shareholder equity
EndBegCap	Altman and Lorris (1976)	The difference between this year's capital and capital additions scaled by the prior year's capital
CapLag	Altman and Lorris (1976)	One period lagged capital
NIATCap	Altman and Lorris (1976)	Net income after taxes scaled by the prior year's capital
NetGross	Altman (1977)	Net operating income scaled by gross operating income
NetTotal	Altman (1977)	Net income scaled by total income
OPTA	Altman (1977)	Total operating expenses scaled by total assets
NWTA	Altman (1977)	Net worth scaled by total assets
PPETA	Altman (1977)	The total value of plants property and equipment scaled by total assets
NOINI	Altman (1977)	Non-operating income scaled by net income
NINW	Altman (1977)	Net income scaled by net worth

Continued on next page

Name	Source	Description
OfficeExp	Altman (1977)	Office building expenses scaled by operating income
ESTA	Altman (1977)	Earned surplus scaled by total assets
PaidEquity	Santomero and Vinso (1977)	Total shareholder equity
TotCap	Santomero and Vinso (1977)	Total capital
CapAssRat	Santomero and Vinso (1977)	Capital scaled by total assets
NISale	Lee et al., (2012)	Net income scaled by total sales
EBITDATA	Lee et al., (2012)	Earnings before interest and taxes, depreciation and amortization scaled by total assets
SETA	Lee et al., (2012)	Total shareholder equity scaled by total assets
SEIATA	Lee et al., (2012)	The difference between shareholder equity and intangible assets scaled by total assets
Lev2	Lee et al., (2012)	Total assets minus intangible assets, cash holdings and the value of land and buildings
CLCTA	Lee et al., (2012)	The difference between current liabilities and cash holdings scaled by total assets
ARSale	Lee et al., (2012)	Accounts receivable scaled by total sales
APSale	Lee et al., (2012)	Accounts payable scaled by total sales
InvGrow	Lee et al., (2012)	Year over year percentage change in inventory
LTGrow	Lee et al., (2012)	Year over year percentage change in total liabilities
CashGrow	Lee et al., (2012)	Year over year percentage change in cash holdings
aftret_eq	WRDS Ratio	After-tax return on average common equity
aftret_equity	WRDS Ratio	After-tax return on total shareholder equity
capital_ratio	WRDS Ratio	Fraction of capital made up by debt
cash_lt	WRDS Ratio	Cash balance scaled by total liabilities
cash_ratio	WRDS Ratio	Cash and cash equivalents scaled by current liabilities
curr_debt	WRDS Ratio	Current liabilities scaled by total liabilities
de_ratio	WRDS Ratio	Total debt scaled by total equity
debt_assets	WRDS Ratio	Debt to assets ratio - different specification to constructed ratio "Leverage" above
debt_at	WRDS Ratio	Alternative specification of the debt to assets ratio
debt_capital	WRDS Ratio	Total debt scaled by total capital
debt_ebitda	WRDS Ratio	Total debt scaled by earnings before interest, taxes, depreciation and amortization
evm	WRDS Ratio	Enterprise value scaled by earnings before interest, taxes, depreciation and amortization
gprof	WRDS Ratio	Alternative specification of profits scaled by assets

Continued on next page

Name	Source	Description
invt_act	WRDS Ratio	Inventory scaled by current assets
lt_debt	WRDS Ratio	Long-term debt scaled by total liabilities
lt_ppent	WRDS Ratio	Total liabilities scaled by total tangible assets
profit_lct	WRDS Ratio	Profits before depreciation scaled by current liabilities
quick_ratio	WRDS Ratio	Acid test ratio
rd_sale	WRDS Ratio	Research and development expenditures scaled by total sales
rect_act	WRDS Ratio	Total receivables scaled by current assets

Appendix D - Factor Zoo Selected Variables

Table 19: Feng, Giglio & Xiu Factor Zoo Selected Variables

Name	Source	Description
AdExp	Chan, Lakonishok and Sougiannis (2001)	Advertising expenditures
BMdec	Fama and French (1992)	Book-to-market using December market equity
ChInvIA	Abarbanell and Bushee (1998)	Change in capital expenditures above or below an industry benchmark
ChNWC	Soliman (2008)	Change in working capital
ChTax	Thomas and Zhang (2011)	Quarterly changes in tax expenses
CompEquIss	Daniel and Titman (2006)	Share issuance for both cash and services (i.e. including employee stock plans)
DivSeason	Hartzmark and Salomon (2013)	Month in which a dividend is expected to be issued
EarningsSurprise	Foster, Olsen and Shevlin (1984)	Sign and magnitude of earnings forecast errors
EarnSupBig	Hou (2007)	Sign and magnitude of earnings forecast errors for large firms
EBM	Penman, Richardson and Tuna (2007)	Net operating assets scaled by price
EP	Basu (1977)	Price earnings ratio
NetDebtPrice	Penman, Richardson and Tuna (2007)	Market value of debt scaled by market value of equity
FirmAge	Barry and Brown (1984)	Months since listed on an exchange
RD	Chan, Lakonishok and Sougiannis (2001)	R&D expenditures scaled by market capitalization
VolMkt	Haugen and Baker (1996)	Trading volume scaled by market equity

Appendix E - Altman Z, WLS Re-Estimation and Random Forest Portfolio Comparisons

Figure 9: Comparison of SSD_e, Altman and WLS Re-Estimation of Altman Cumulative Returns

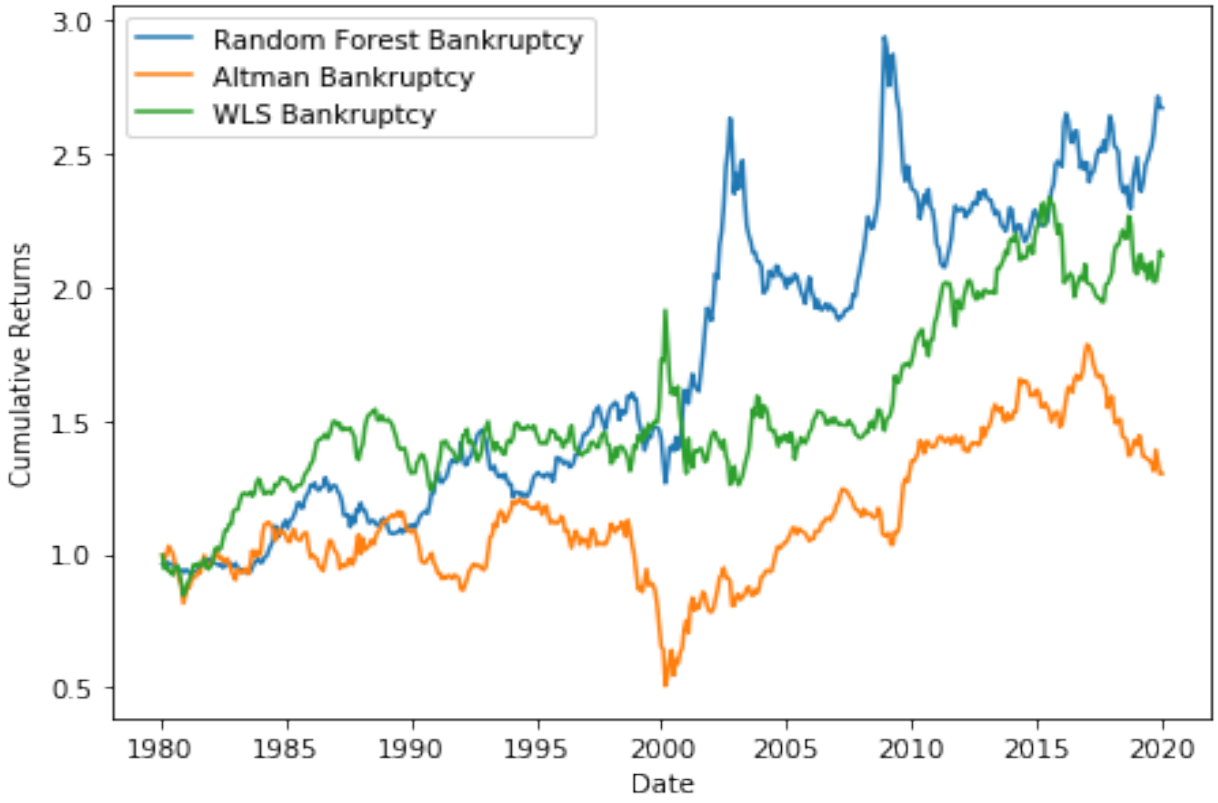


Figure 10: Comparison of SSD_e, Altman and WLS Re-Estimation of Altman Cumulative Returns

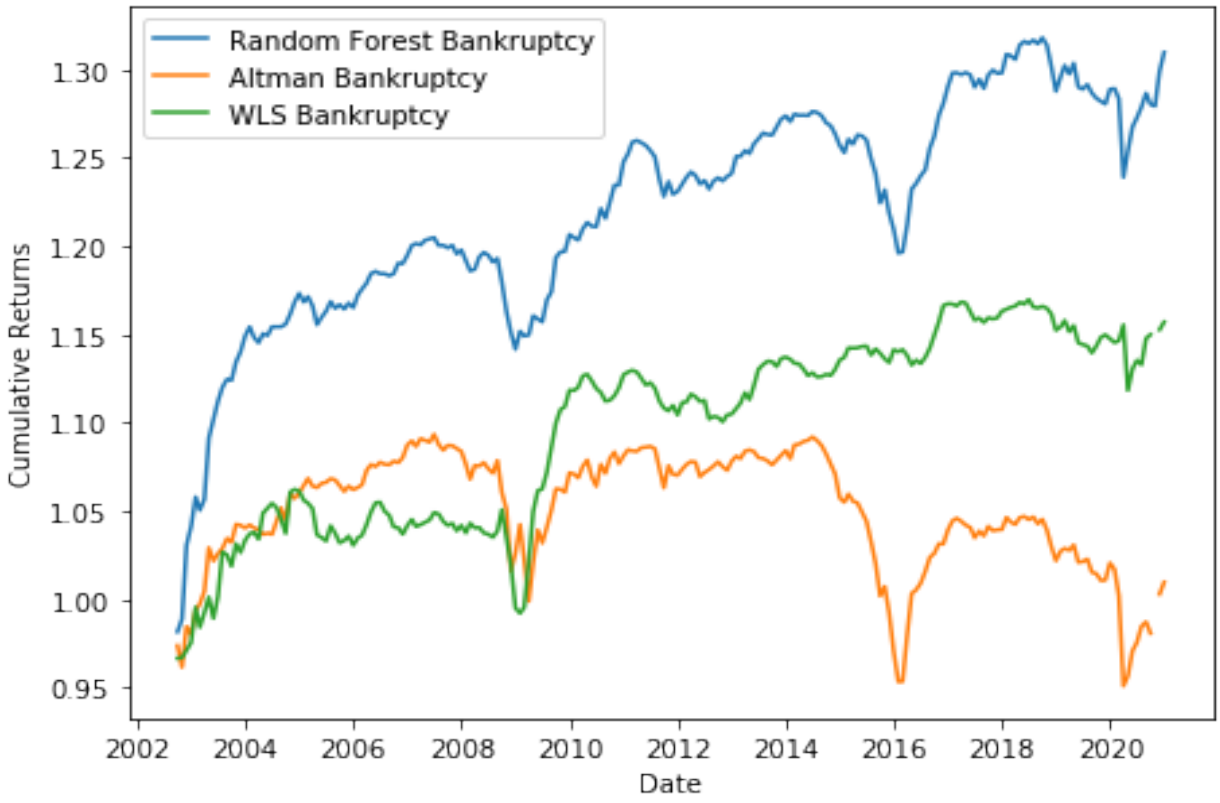


Figure 11: Comparison of SSD_e, Altman and WLS Re-Estimation of Altman Cumulative Returns

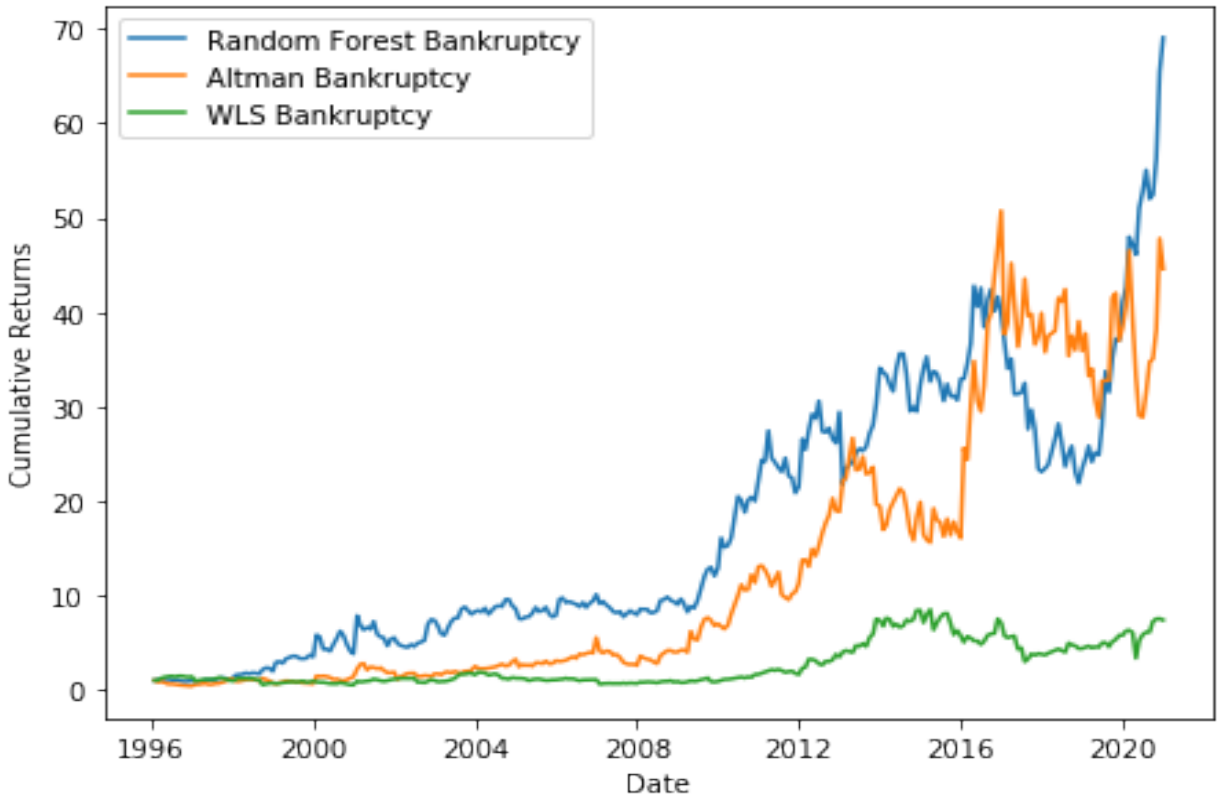


Figure 12: Comparison of SSD_e, Altman and WLS Re-Estimation of Altman Cumulative Returns