HEC MONTRÉAL

Improvement of the credit risk stress testing methodology for corporate bonds using machine learning with covariate shift adaptation

par

Robert Timper

Sciences de la gestion
(Option Applied Financial Economics)

Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences
(M. Sc.)

January 2021

Abstract

Credit stress testing has become an important risk management practice in the financial industry and is required for financial institutions by the regulator. It plays a crucial role in monitoring a portfolio's resilience against severe macroeconomic shocks. The challenges in constructing a model for these stress tests lie in the quantitative representation of the credit risk that is related to the macroeconomic variables as well as the model accuracy in times of stressed economic conditions. I have chosen a one-parameter model based on rating transitions to represent the systematic credit risk. To estimate this credit risk using macroeconomic variables, I will present a machine learning model, taking into account a broad range of macroeconomic variables, that is able to explore non-linearities. Additionally, I have implemented a covariate shift adaptation, to further increase the model accuracy for times of economic stress. It will be shown that the model accuracy increases by roughly 50% with the machine learning model compared to a linear benchmark model.

Keywords

Credit Risk, Risk Management, Stress Testing, Z-Factor, Machine Learning, SVR, Covariate Shift, Lasso Regression

Résumé

Les tests de résistance du risque de crédit sont devenus une pratique importante de gestion des risques dans le secteur financier et sont exigés des institutions financières par le régulateur. Ils jouent un rôle crucial dans le suivi de la résilience d'un portefeuille face aux chocs macroéconomiques sévères. Les défis de la construction d'un modèle pour ces tests de résistance résident dans la représentation quantitative du risque de crédit liée aux variables macroéconomiques ainsi que dans la précision du modèle en périodes de conditions économiques difficiles. J'ai choisi un modèle à un facteur basé sur des transitions de notations pour représenter le risque de crédit systématique. Pour estimer ce risque de crédit à l'aide de variables macroéconomiques, je présente un modèle d'apprentissage automatique, prenant en compte un large éventail de variables macroéconomiques, capables d'explorer les non-linéarités. De plus, j'ai implémenté une adaptation de décalage de covariables, pour augmenter encore plus la précision du modèle pour les périodes de stress économique. Il est démontré que la précision du modèle augmente d'environ 50% avec le modèle d'apprentissage automatique par rapport à un modèle de référence linéaire.

Mots-clés

Risque de crédit, gestion des risques, tests de résistance, facteur Z, apprentissage automatique, SVR, changement de covariable, régression par lasso

Contents

List of Tables

List of Figures

List of acronyms

ACF Autocorrelation Function

ADF Augmented Dickey-Fuller

AIRB Advanced Internal Rating-Based

AIWSVR Adaptive Importance-Weighted Support Vector Regression

ARIMA Autoregressive Integrated Moving Average

BFGS Broyden-Fletcher-Goldfarb-Shanno

BIS Bank of International Settlement

BLUE Best Linear Unbiased Estimator

CB Central Bank

CCAR Comprehensive Capital Analysis and Review

CV Cross-Validation

DBTS Domestic Bank Tightening Standards

DSR Debt-Service-Ratio

GDP Gross Domestic Product

FRED Federal Reserve Economic Data

HEC Hautes études commerciales

IFRS International Financial Reporting Standards

IMF International Monetary Fund

ISM Institute for Supply Management

KPSS Kwiatkowski–Phillips–Schmidt–Shin

MAE Mean Absolute Error

MBS Mortgage Backed Securities

MSc Maîtrise

MSE Mean Squared Error

NIPA National Income and Product Account

NN Neural Network

NIPA National Income and Product Accounts

NPISH Non-Profit Institutions Servings Households

OLS Ordinary Least Square

PACF Partial Autocorrelation Function

PI Price Index

PIT Point-In-Time

PMI Purchasing Managers' Index

PNFS Private Non-Financial Sector

PP Phillips-Perron

QD Quarterly Difference

RBF Radial Basis Function

RF Random Forest

Repo Repurchase Agreement

SFA Single Factor Analysis

S&P Standard & Poor's

SVM Support Vector Machine

SVR Support Vector Regression

TTC Through-The-Cylce

VIF Variance Inflation Factor

VIX CBOE Volatility Index

WTI West Texas Intermediate

YD Yearly Difference

Acknowledgements

I would like to thank my thesis director, M. Georges Dionne, for supporting me throughout this research project. He was always available and very quick to share his extensive knowledge, provide me with valuable feedback and ideas that were essential for the success of this thesis. I would also like to express my gratitude towards the manager and the team at the international financial institution that allowed me to use their data for this research. Their support and feedback was invaluable and I want to thank them for giving me this opportunity.

Further, I would like to thank my family and friends who have supported me during this work. I want to especially thank Mariana, for being there for me every day and always encouraging me. It would not have been possible without you.

Last but not least, I thank the jurors who have accepted to evaluate my work.

Introduction

The importance of stress testing has increased dramatically ever since the Great Financial Crisis and is an important pillar of the regulatory framework for financials institutions today. In the US, this is known as Comprehensive Capital Analysis and Review (CCAR) which is required for bank holding corporations since the 2010 Dodd-Frank-Act and overseen by the Fed. These stress test play an important role in assessing the sufficiency of banks' capital requirement when facing severe economic shocks. The stress tests are applied to all of a bank's holdings, including corporate credit portfolios which I will focus on.

The main challenge in creating a stress testing model arises in accurately linking the macroeconomic environment to the creditworthiness of the corporate credit portfolio. To quantify the systematic credit risk, I will implement the framework of Belkin et al. (1998b) which relies on a conditional transition matrix approach. The transition matrices are constructed with corporate credit rating data. I will use the internal rating data for a corporate credit portfolio of an international financial institution. Based on this portfolio, with a 10-year history of quarterly ratings, an indicator for systematic credit risk can be obtained. This so-called Z-Factor can then be linked to the macroeconomic environment as in Bangia et al. (2002). However, it is quite challenging to precisely model this relationship.

Traditionally, the systematic credit risk is estimated with a linear regression and classical macroeconomic variables such as GDP growth, unemployment rate and interest rates. In this master thesis, I will apply a machine learning model for the estimation of the Z-

Factor and extend the macroeconomic variable candidates. The use of a machine learning model aims to explore non-linearities in the relationship between the systematic credit risk and the macroeconomic environment. The extended scope of macroeconomic variables intends to improve the accuracy when modeling this relationship, by providing information not previously considered. Further, I will introduce the concept of covariate shift, to improve the model accuracy for periods of adverse economic conditions. A model with covariate shift adaptation can take into account a different distribution of the independent variables for new observations. This is clearly of particular interest when applying a stress scenario. The results of this approach will be compared to a linear Lasso regression. The Lasso regression has previously been successfully applied in the context of credit stress testing as Chan-Lau (2017) shows.

The remainder of the thesis is organized as follows. The next chapter provides an overview of the existing literature regarding credit stress testing and machine learning application in this field. Chapter 2 describes the economic and rating data that is used in the thesis. In chapter 3 the model and methodology are presented. Chapter 4 reports the empirical results of the various models applied to the data. In Chapter 5 these results are discussed and interpreted, followed by a brief conclusion and future research topics.

Chapter 1

Literature Review

Credit stress testing models allow financial institutions to simulate economic events and evaluate their impact on their credit portfolio. Since the Great Financial Crisis, the majority of central banks, including the Fed, have regulatory requirements for stress testing of financial institutions. Stress testing of corporate credit portfolios is also included in the Basel III framework. Therefore, the models used for those stress tests have become more important and have drawn researchers' interests as well. Two model types exist that allow the analysis of credit risk for the stress testing purpose. One is a reduced form model where a credit event is triggered by an exogenous macroeconomic shock, the other is a structural model that is based on firm specific metrics. A commonly used reduced form model for the credit stress testing methodology in financial institutions is the one-parameter model. This model quantifies a portfolio's systematic credit risk in a single number, based on the credit migrations. It is founded on the framework developed by Vasicek (1987) who demonstrated that the credit risk of a firm's debt can be split into an idiosyncratic and a systematic component.

$$z = bx + a\sum \varepsilon_i \tag{1.1}$$

Equation 1.1 from Vasicek (1987) displays this decomposition with x as the systematic component and ε_i as the idiosyncratic component. This framework is based on Merton (1974) who describes the underlying asset value as a geometric Brownian motion over

time which leads to the debt holders actually holding a put option on the firm's value. Belkin et al. (1998b) adopt this framework to extract a single synthetic credit indicator from credit rating transitions by decomposing the change of creditworthiness, X, in the same two components as Vasicek (1987).

$$X = \sqrt{1-\rho}Y + \sqrt{\rho}Z \tag{1.2}$$

With Y representing the idiosyncratic component, Z representing the systematic component and ρ as the correlation coefficient between Z and X, as seen in equation 1.2. This is founded on the assumption that the credit risk has a standard normal distribution, which implies the the systematic credit risk, Z-Factor, and the idiosyncratic credit risk are standard normal distributed as well. As Bandt et al. (2013) point out, this assumption is appropriate and can be relaxed for the Z-Factor. Based on this framework, Belkin et al. (1998b) determine the systematic credit risk component as the distance between the long-term average transition matrix and the transition matrix at any point in time. The long-term average transition matrix is the so called through-the-cycle matrix. For this matrix, the continuous time duration method is to be preferred over the discrete-time cohort method as Schuermann and Hanson (2004) show. They conclude that the cohort method can be inaccurate and produce inefficiencies, both in statistical and economic terms. This is caused by two characteristics of credit rating data, observed over a limited period. It is unknown what happened to firms before or after the observed period. This is unaccounted for in the cohort method. Further, the duration method attributes positive but small probabilities to all transitions even when no such transition is observed, whereas the cohort method would assign a zero probability to those transitions. This is important as credit rating data often lacks observed transitions, especially, form high ratings to low ratings or default, but it is safe to say those probabilities are not zero in reality. The duration method is based on a Markov process which assumes time-homogeneity. This is not necessarily given, specifically with downgrade movements as Lando and Skødeberg (2002) point out. Schuermann and Hanson (2004), however, prove that the impact between a parametric

duration method, assuming time homogeneity, and a non-parametric approach, relaxing the time homogeneity assumption, is marginal.

When constructing rating transition matrices, it is important to account for the different characteristics of the obligors, to obtain stable rating transition probabilities, as Nickell et al. (2000) demonstrate. Credit rating distributions vary across obligor origin and industry. They find significant variations between financial and industrial obligors as well as between US and non-US entities. These results are confirmed by Kadam and Lenk (2008) who find significant differences for the transition matrices between the financial, industrial and utility sector. Therefore, the portfolio of obligors in this work consists only of US entities and only industrial obligors are considered, with financial and energy obligors filtered out.

Once a systematic credit risk indicator, the Z-Factor, is obtained from the transition matrices, utilizing the framework of Belkin et al. (1998b), the construction of the actual stress testing model becomes the focus. The challenge here is accurately linking the systematic credit risk to the macroeconomic environment. Figlewski et al. (2012) use a reduced-form Cox hazard model, to estimate rating transition intensities. They find that macroeconomic variables are highly significant in the estimation and their incorporation leads to an increase in the explanatory power. They employ a total of 14 macroeconomic variables, grouped in three categories: general macroeconomic conditions, direction of the economy and financial market conditions. In order to reduce the number of variables and only select the most important ones for their final model, they apply backward selection as a feature selection method. Further, they emphasize the importance of lags, as these economic variables mostly don't have an instantaneous impact on the credit rating transitions. To avoid a further expansion of the number of variables, Figlewski et al. (2012) implement a lag structure. They discover that the general macroeconomic conditions, such as unemployment rate, NBER recession indicator and inflation, have a much smaller impact on the rating transitions than the other categories. This is not an adequate model to conduct credit stress testing but proves the existence of a relationship between macroeconomic conditions and credit rating transitions.

An early application of the Z-Factor in a credit stress test can be found in the work of Bangia et al. (2002). They only use the NBER recession indicator as a representation of the macroeconomic conditions and introduce a regime switching mechanism to decide between expansion and recession transition matrices. Using a Monte Carlo simulation, they achieve a high in-sample accuracy for the default rate and rating distribution. Bandt et al. (2013) expand the range of macroeconomic variables considered and present a model for credit stress testing, applying an OLS regression approach for the model construction. They use the variables GDP growth, unemployment rate, inflation and a 10-year over 3-month yield spread as well as an autoregressive component. However, they do not directly estimate the Z-Factor but use the S&P corporate annual default rate as an intermediary between macroeconomic variables and Z-Factor. This is an unnecessary step and only distorts an already approximated indicator, so I will directly estimate the Z-Factor. GDP growth and inflation are statistically significant in all their models, whereas the unemployment rate and yield spread don't show significance. This confirms the results of Figlewski et al. (2012) in terms of the unemployment rate being insignificant. However, Figlewski et al. (2012) find inflation to be insignificant and yield spread significant but they also use a lot more covariates which can explain these different findings. It is interesting to observe that the autoregressive component is significant and models that leave it out display low Durbin-Watson statistics which is evidence for autocorrelation in the error term. This suggests credit rating transitions display autoregressive behavior.

Machine learning has become increasingly more popular in recent years and its use has expended to the field of risk management as well. Leo et al. (2019) presents a very good overview of the existing applications that include all aspects of risk management and often offer an improvement over traditional methods. Prominent applications of machine learning in risk management feature credit scoring and predicting the probability of default. Machine learning has been successfully applied to classify corporate credit rating as in Lee (2007) or Huang et al. (2004). One of the advantages machine learning presents, is the ability to explore the non-linear relationships that are common in credit risk. The support vector machine algorithm is proven to be very successful in determining credit

risk. Lee (2007) shows that a support vector machine achieves the highest accuracy in predicting credit ratings and is achieving this, despite a relatively small dataset to train the model. Hajek and Michalak (2013) apply a variety of machine learning models to the same classification problem. They try an extensive list of firm specific explanatory variables for the credit scoring and highlight the importance of feature selection prior to the classification problem, as it is shown to improve the model accuracy. The prediction of credit ratings, however, is a classification problem and not a regression problem, like the Z-Factor estimation. Yao et al. (2015) use a support vector regression to estimate the loss given default with accounting and macroeconomic variables. The macroeconomic variables include GDP, unemployment rate, S&P 500 return and treasury bill rate. They find the support vector regression to be more accurate than a linear regression, fractional response regression and a two-stage method. This demonstrates that the relationship between credit risk and economic factors can be described with high accuracy using support vector machines. In Jacobs Jr (2018) a machine learning model is applied to the credit stress testing methodology and shows an improvement over a traditional VAR model in terms of model performance due to its ability to reflect non-linearities. Jacobs Jr (2018) uses the macroeconomic variables, proposed by the Fed stress testing framework, with a 74 quarter history. The variables are transformed, and only stationary ones are considered. Further, variable coefficients have to match their economically intuitive sign. The stress test is applied to three different portfolios: Commercial real estate, consumer credit and commercial & industrial. The later most closely resembles the corporate portfolio used in this thesis. The target variables in this model is the credit loss rate and not a systematic credit risk indicator. The best model for the commercial & industrial portfolio consists of the variables real GDP growth and BBB spread.

As mentioned in Leo et al. (2019), Lasso regressions present a good modeling approach to obtain sparse and approximately unbiased results for the relationship between macroeconomic variables and credit losses. Chan-Lau (2017) presents the application of a Lasso regression to a credit stress testing framework where the probability of default is predicted. The dataset consists of median probabilities of default for ten industrial sectors

in an advanced emerging market economy. The estimations are done separately for each of the sectors, using 13 macroeconomic variables. The variables including exchange rates, interest rates, GDP growth and unemployment rate. The data is on a quarterly bases and includes 96 observations with up to four lags included for each variable. The advantage of the Lasso regression over OLS is the ability to handle high dimensional datasets. The work by Chan-Lau (2017) shows that a Lasso model is adequate as a benchmark model.

Chapter 2

Data

This chapter will describe the data that has been used in this thesis. Two different datasets were required: Credit rating data from a portfolio of corporate bonds to calculate the Z-Factor and macroeconomic data to build a model for the estimation of that Z-Factor.

2.1 Credit Rating Data

The credit rating data was obtained from the corporate bond portfolio of a large international financial institution. The portfolio solely includes U.S. entities. In the scope of this thesis, only the corporate segment of the portfolio is considered which excludes financial institutions, and oil and gas companies. The dataset consists of 6035 obligors which are rated on a monthly basis. When analyzing rating transitions, it is common to analyze these over smaller frequencies than monthly. This is due to the fact that rating changes tend to happen gradually, as ratings are commonly reviewed yearly and therefore most studies of rating transitions look at these transitions on a quarterly or yearly basis. For the majority of this thesis, I will consider ratings on a quarterly basis, but I will test the model performance with monthly frequency. The observed ratings range from October 2007 until June 2020 which leads to 51 quarters or 153 months of observations. For the quarterly ratings, the last monthly rating of each quarter is taken. This leads to 52,866 quarterly ratings and 131,725 monthly ratings.

External Ratings		Internal Ratings	
Moody's	S&P	Rating	Rating Bucket
Aaa	AAA	1	1
Aa1	AA+	2+	
Aa2	AA	2	2
Aa3	AA-	2-	
A1	A+	3+	
A2	A	3	3
A3	A-	3-	
Baa1	BBB+	4+	
Baa2	BBB	4	4
Baa3	BBB-	4-	
Ba1	BB+	5+	
Ba2	BB	5	5
Ba3	BB-	5-	
B1	B+	6+	
B2	B	6	6
B3	b-	6-	
Caa1	CCC+	7+	
Caa2	CCC	7	7
Caa3	CCC-	7-	
Default	Default	8	
Default	Default	9	8
Default	Default	10	

Table 2.1: Rating Overview

In the process of preparing the data it became apparent that some obligors have a gap in their rating history of 1 or 2 month, caused by ratings not being updated in time. To create a more complete dataset, these ratings were filled according to the more conservative out of the rating before and after the gap. This leads to 4,008 quarterly and 8,212 monthly ratings getting filled. Ratings in the data range from 1-20 with 1 being the best rating. These correspond to the typical rating agency ratings as shown in Table 2.1. To reduce noise and increase the number of observations per rating, the ratings are grouped into 8 buckets. The buckets are formed with the 3 ratings that correspond to one letter of the typical agency rating scale. The 8th bucket is the default bucket and contains companies that have defaulted. A default is defined as a counterparty that under International Financial Reporting Standards (IFRS) accounting standards is determined impaired. The

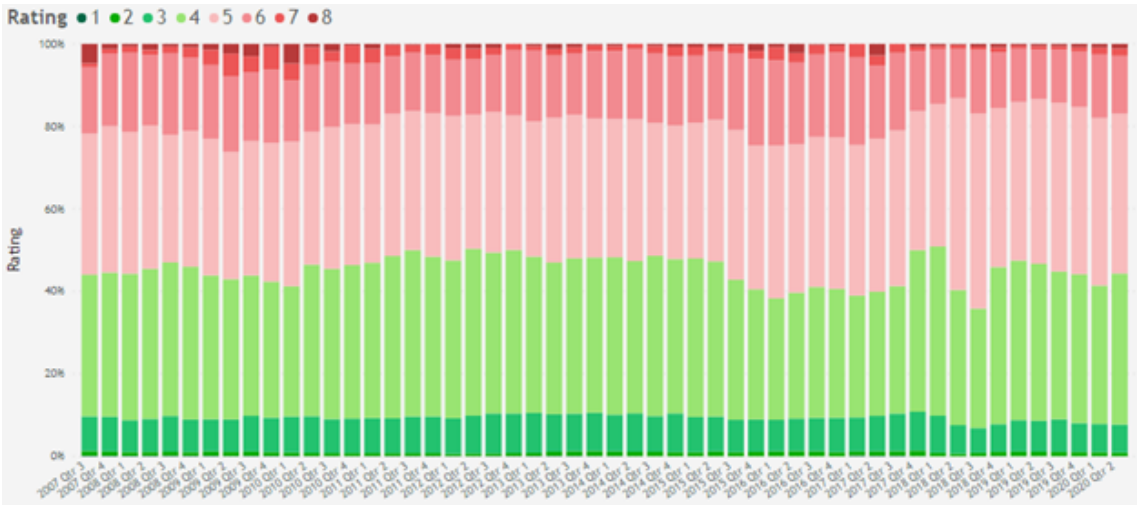


Figure 2.1: Rating per Quarter

development of the quarterly ratings over time can be observed in Figure 2.1.

2.1.1 Sectors

The corporate segment of companies can further be divided into more granular sections. An overview of the sections can be found in Table 2.2. Since the number of obligors are small and to reduce the number of sections, some sections are combined into one for the purpose of this work. The sections particular services and business services are combined to a service section, the construction and real estate sections are coupled to form a housing section and finally the auto industry and multi activity group sections form a miscellaneous section.

2.2 Economic Data

Since the Z-Factor is a systematic factor it can be estimated using macroeconomic variables as pointed out by Bangia et al. (2002). Typically only a few classical macroeconomic variables are considered for the estimation of the Z-Factor in other research. The Fed proposes a variety of variables in the CCAR framework which I will use as a basis. However, I will extend the range of macroeconomic variables and indicators further,

Section	Obligors
Agriculture	376
Consumption Goods	326
Capital Goods	589
Intermediate Goods	639
Utilities	486
Commerce	1035
Transportation	402
Media & Telecommunications	264
Real Estate	499
Services	1262
Miscellaneous	310

Table 2.2: Section Overview

specifically to credit related ones. Since the portfolio only concerns U.S. entities, all variables are specific to the U.S. Not all data is available on a monthly basis and therefore the monthly estimation will be based on a smaller set of variables.

2.2.1 Economic Fundamentals

Income is a standard economic variable and offers a broad indicator of the state of the economy which affects credit quality. The real and nominal GDP growth rate as well as the real disposable income growth rate in the U.S. are considered as a proxy for income. This data is officially published on a quarterly basis but IHS Markit provides monthly nominal and real GDP calculations that resemble the official ones. The IHS Markit data is used for the monthly and quarterly estimation to avoid differences due to different sources. The real disposable income growth is sourced from the FRED database and only available on a quarterly basis.

To consider the labor market effects on the economy and the consequences this has on credit quality, the unemployment rate, initial jobless claims and continuous jobless claims

are sourced from the FRED database. The seasonally adjusted series are chosen for all three variables. While the unemployment rate is published on a monthly basis, the jobless claims statistics are weekly, and the average is taken to obtain the monthly or quarterly value, respectively.

To account for the effect, the real estate sector can have on the economy and credit quality, a commercial real estate and housing price index are taken into consideration. The commercial real estate price index from the FRED database is selected as well as the house price index published by the federal housing finance agency. The commercial real estate price index is only available on a quarterly basis.

Another economic variable considered is the U.S. export goods volume from the direction of trade statistic, published monthly by the IMF.

2.2.2 Financial Market Data

Financial market data such as stock price indices can be good indicators for the economy and interest rates impact credit quality through their pricing power on debt. Two stock market indices are utilized here: The Dow Jones total stock market index from the FRED database and the S&P 500 industrial index where the last price for the respected month or quarter is observed.

Four interest rates are considered and used to calculate: The BBB corporate bond spread over the 10-year US treasury yield, the 10-year US treasury yield over the 3-month US T-bill and the 5-year US treasury over the 3-month US T-bill. This is done using the monthly and quarterly average rates, respectively. All rates are constant maturity rates.

The market volatility index (VIX) is also considered as it indicates periods of stress in financial markets which coincide with periods of stressed credit conditions. The US dollar index is used to represent the impact of US dollar weakness or strength. Finally, the WTI Crude Oil Price is applied as a proxy for commodity prices.

2.2.3 Government Data

The recent decades have seen financial markets and the economy being increasingly affected by government intervention through fiscal spending or central bank policy. To capture this effect, fiscal and federal reserve balance sheet data is considered. The US treasury general account at the federal reserve which indicates future fiscal spending is used to capture government spending.

The impact of federal reserve policy is difficult to quantify, as often the announcement of policies and the forward guidance have a bigger effect, for example on yields, than the programs themselves. For this reason, several different positions on the central bank balance sheet are tested as variables. The treasury securities and mortgage backed securities that the federal reserve purchased on the secondary market are each used as an input variable. These two holdings combine for the total securities held by the federal reserve which is also used as a variable. Repos with commercial banks which occur mostly in times of financial distress for banks are considered. Additionally, liquidity swaps with other central banks, which also occur mostly in times of financial distress, are used as a variable. Loans given out by the federal reserve directly to companies are also considered. Finally, the total assets on the balance sheet are taken as a variable. Further, the M2 money supply and M2 velocity are considered as measures that capture fiscal and central bank spending. M2 velocity is only published quarterly.

2.2.4 Corporate Balance Sheet Data

To incorporate in detail the profitability of companies and with that their ability to repay their debt, the profit of all corporations and more specific non-financial corporations is obtained form the National Income and Product Account (NIPA). Additionally, the debt burden is measured using the debt service ratio of non-financial corporations, the private non-financial sector and households, and NPISH from the bank of international settlement (BIS). All of this data is only available quarterly.

2.2.5 Survey Data

Surveys can often be used as good leading indicators compared to GDP which is a lagging variable, as it is only calculated and published after the fact. A popular survey is the Institute for Supply Management's purchasing manager index (PMI). In this work, the manufacturing and services PMIs are considered. Further, the federal reserve publishes a senior loan officer survey among bankers. From this survey, the lending tightening standards for commercial and industrial loans to large and middle market firms, and for consumer loans and credit cards are used as indicators for tighter credit conditions. These two lending tightening surveys are only available quarterly.

An overview of all the variables with their expected coefficient in the Z-Factor estimation, based on economic theory, can be found in Table 2.3.

Variable	Quarterly	Monthly	Format	Expected Coefficient
Real GDP Growth	✓	✓	rate	+
Nominal GDP Growth	✓	✓	rate	+
Real Disposable Income Growth	✓	✓	rate	+
Unemployment Rate	✓	✓	rate	-
Continuous Jobless Claims	✓	✓	level	-
Initial Jobless Claims	✓	✓	level	-
Commercial Real Estate Price Index	✓	x	level	+
House Price Index	✓	✓	level	+
US Export Volume Goods	✓	✓	level	+
Recession Dummy	✓	✓	binary	-
Dow Jones Total Stock Market Index	✓	✓	level	+
S&P Industrial	✓	✓	level	+
BBB Spread	✓	✓	rate	-
10 year - 3 month Spread	✓	✓	rate	-
5 year - 3 month Spread	✓	✓	rate	-
VIX	✓	✓	level	-
WTI Oil Price	✓	✓	level	-
M2	✓	✓	level	-
M2 Velocity	✓	✓	level	+
Treasury Securities	✓	✓	level	-
MBS	✓	✓	level	-
Total Securities	✓	✓	level	-
Repos	✓	✓	level	-
Liquidity Swaps with other central bank	✓	✓	level	-
Loans	✓	✓	level	-
Total Assets	✓	✓	level	-
US Treasury General Account	✓	✓	level	-
US Dollar Index	✓	✓	level	-
NIPA Profitability	✓	x	level	+
NIPA non-financial Profitability	✓	x	level	+
DSR (non-financial Corporate)	✓	x	level	-
DSR (PNFS)	✓	x	level	-
DSR (PNFS, Households &NPISH	✓	x	level	-
ISM Manufacturing PMI	✓	✓	level	+
ISM Services PMI	✓	✓	level	+
DBTS for Commercial and Industrial Loans	✓	x	level	-
DBTS on Consumer Loans and Credit Cards	✓	x	level	-

Table 2.3: Variable Overview

Chapter 3

Methodology

This chapter will present the methodology applied to calculate the Z-Factor from the portfolio of corporate bonds and then use these to build a model that accurately predict future Z-Factors. The calculation of the Z-Factor is based on the methodology, presented by Vasicek (1987), which represents a simple one-parameter model to describe the credit risk of a portfolio. It is based on the Merton (1974) framework, with the Z-Factor as a proxy for the change in the underlying asset value. Vasicek (1987) defines the portfolio default rate as a function of the correlation of asset values, the single firm default probability and a systematic risk factor. This framework is widely used today in banks' economic capital models, credit stress testing models and the Basel framework for Advanced Internal Rating-Based (AIRB) regulatory credit risk capital. In the one-parameter model, the continuous normally distributed credit indicator X, can be split into an idiosyncratic component, Y, and a systematic component, Z, as shown by Belkin et al. (1998b) These components form X using the correlation coefficient, ρ, between Z and X:

$$X = \sqrt{1-\rho}Y + \sqrt{\rho}Z \tag{3.1}$$

Belkin et al. (1998b) then go on to define the Z-Factor as the deviation in the transition of ratings from a long-term historical average transition of ratings.

3.1 Z-Factor calculation

Since the Z-Factor represents the deviation of rating transitions from a historical average of rating transitions, it is necessary to calculate a historical transition matrix and matrices for every quarterly observation of ratings. In this work, I will construct a Trough-The-Cycle (TTC) transition matrix as the historical average, and I will construct Point-In-Time (PIT) transition matrices for every observation date. By calculating the deviation between the respective PIT matrix and the TTC matrix at every observation, the Z-Factors will be obtained. There are two common methods to construct transition matrices: the cohort method and the duration method. The transition probabilities in the cohort method are obtained, by observing the changes of ratings in each bucket from one period to another. For the duration method, the rating changes are compared to the time an obligor remains in a certain rating bucket.

3.1.1 Through-The-Cycle transition matrix

The TTC matrix represents the average transition matrix over the observed time frame. It is generated using the duration method. The duration method is chosen here because unlike the cohort method it gives small but non-zero probabilities to transitions, even if there are no observations for such a transition. In a risk context, it is important to capture even such rare events that are not represented in the data but can possible be realized. For the duration method, the rating changes over the whole timeframe are counted and then divided by the amount of time spent in each rating state, to obtain a matrix of transition intensities. This matrix is also called the generator matrix and is assumed to be time homogenous. With this assumption, the generator matrix can be assumed to follow a Markov process. Given that the generator matrix follows a Markov process, the following formula proposed by Lando and Skødeberg (2002) can be used to calculate the transition matrix, $P(t)$, using the generator matrix by applying the matrix exponential function:

$$P(t) = P(0,t) = exp(\Lambda t) = \sum_{k=0}^{\infty} \frac{\Lambda^k t^k}{k!} = I + t\Lambda + \frac{(t\Lambda)^2}{2!} + ... \tag{3.2}$$

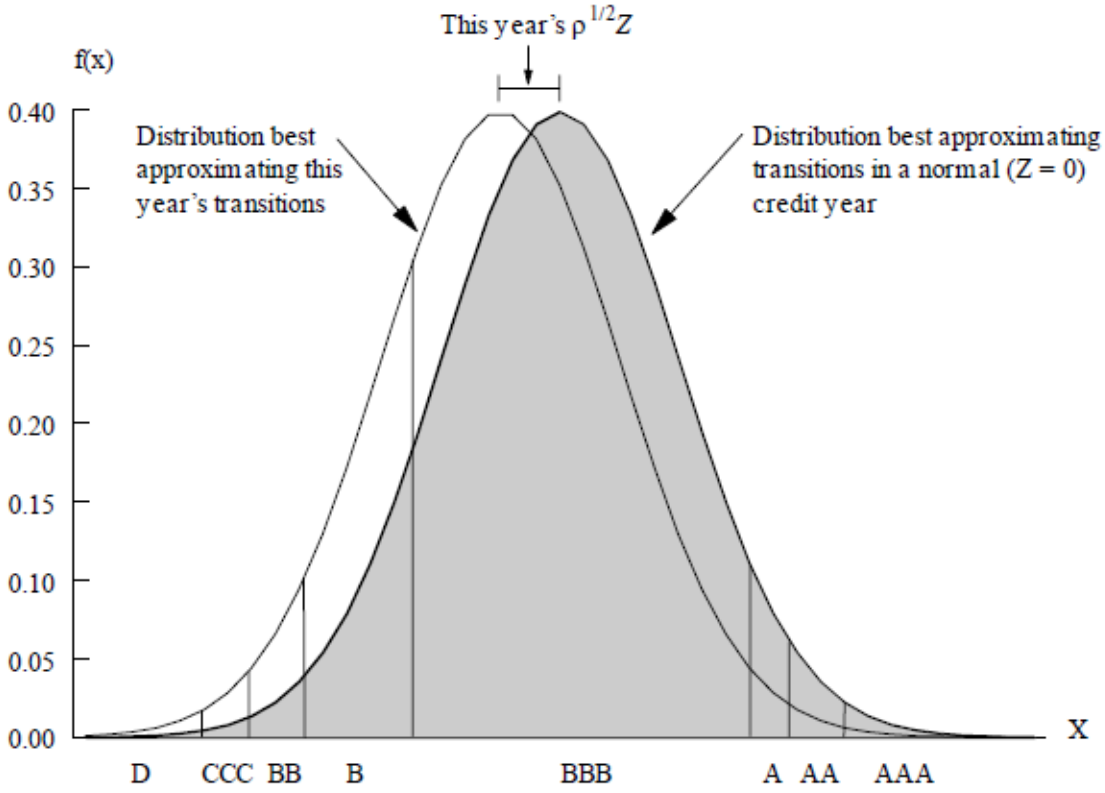


Figure 3.1: Graphical presentation of the Z-Factor as in Belkin et al. (1998b)

In equation 3.2, Λ is the generator matrix, t is the timeframe and k is the rating bucket. Table 3.1 shows the obtained TTC matrix.

3.1.2 Point-In-Time Transition Matrices

The PITs are calculated between every two consecutive observation dates. Consequently, there is one transition per obligor and therefore, the time component captured in the duration method can be disregarded here. It is therefore logical, to apply the cohort method. The transition probabilities with the cohort method are computed as the number of ratings per bucket at the end of the period as a share of the total number of ratings in the bucket at the start of the period. This leads to the probabilities in each row in the transition matrix adding up to 100%.

Initial Rating	End-of-period Rating							
	1	2	3	4	5	6	7	8
1	95.28%	3.60%	0.08%	1.04%	0.01%	0.00%	0.00%	0.00%
2	0.00%	95.94%	3.92%	0.14%	0.00%	0.00%	0.00%	0.00%
3	0.02%	0.26%	97.28%	2.35%	0.08%	0.00%	0.00%	0.02%
4	0.00%	0.03%	0.46%	97.80%	1.57%	0.06%	0.04%	0.03%
5	0.00%	0.02%	0.03%	1.36%	96.57%	1.66%	0.09%	0.26%
6	0.00%	0.00%	0.05%	0.11%	2.57%	94.71%	1.38%	1.18%
7	0.00%	0.00%	0.14%	0.01%	0.87%	4.65%	84.89%	9.44%
8	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%

Table 3.1: TTC Matrix

3.1.3 Z-Factor Extraction

Having calculated the TTC and PITs, the Z-Factor can be calculated as the deviation in the rating distribution between the PIT and TTC at every observation as visualized in Figure 3.1. First, bins are defined for each rating transition according to the normal distribution. The borders of the bins are defined as the difference between the Z-scores of the normal distribution of the end-of-period and initial transition probabilities:

$$P(G,g) = \Phi(R^{Eop}) - \Phi(R^{init}) \tag{3.3}$$

This leads to the bins for the TTC matrix shown in Table 3.2.

These rating transitions still represent the X, the credit indicator. In the next step, the decomposition of this indicator into the systematic (Z) and idiosyncratic component(Y) takes place.

$$X = \sqrt{1-\rho}Y + \sqrt{\rho} * Z \tag{3.4}$$

Since the portfolio consists of a large number of obligors, the idiosyncratic component (Y) can be assumed to be eliminated through diversification as Belkin et al. (1998a) argue and thus, Z is a sufficient estimate to determine X. Now, a value for Z can be found, so that the borders of the bins are best approximating the transition probabilities of every PIT. To determine the difference, the fitted transition probabilities need to be calculated first. They are defined as follows:

$$\Delta(R^{Eop}, R^{init}, Z_t) = \Phi(\frac{R^{Eop} - \sqrt{\rho}Z_t}{\sqrt{1-\rho}}) - \Phi(\frac{R^{init} - \sqrt{\rho}Z_t}{\sqrt{1-\rho}}) \tag{3.5}$$

For every Z, the difference between the long-term TTC and the fitted PIT transition probabilities is minimized with a negative maximum likelihood estimation instead of the weighted mean squared difference as Belkin et al. (1998b) use:

$$L(\rho, Z_1, ..., Z_T) = \prod_{i=1}^{7}\prod_{j=1}^{8}\prod_{t=1}^{T}(M_{i,j}^{Z_t,\rho})^{n_{i,j,t}} \tag{3.6}$$

where $n_{i,j,t}$ represents the number of observed migrations from rating i to rating j between t and $t+1$.

Initial Rating	End-of-period Rating							
	1	2	3	4	5	6	7	8
1	[∞ , -1.67]	[-1.67 ,-2.28]	[-2.28 ,-2.31]	[-2.31,-3.74]	[-3.74 ,-4.34]	[-4.34 ,-4.47]	[-4.47, -4.65]	[-4.65,−∞]
2	[∞ , 4.52]	[4.52 ,-1.74]	[-1.74, -2.99]	[-2.99,-4.04]	[-4.04 , -4.46]	[-4.46 ,-4.49]	[-4.49, -4.50]	[-4.50 , −∞]
3	[∞ , 3.61]	[3.61 ,2.78]	[2.78 , -1.97]	[-1.97,-3.10]	[-3.10 , -3.57]	[-3.57, -3.59]	[-3.59,-3.60]	[-3.60,−∞]
4	[∞ , 3.91]	[3.91 ,3.37]	[3.37 , 2.58]	[2.58, -2.12]	[-2.12 , -3.00]	[-3.00 ,-3.19]	[-3.19 ,-3.41]	[-3.41, -∞]
5	[∞ , 4.97]	[4.97 ,3.51]	[3.51,3.27]	[3.27 , 2.19]	[2.19 ,-2.05]	[-2.05,-2.70]	[-2.70 ,-2.80]	[-2.80 , -∞]
6	[∞ , 5.29]	[5.29 ,4.48]	[4.48 ,3.31]	[3.31,2.95]	[2.95 ,1.92]	[1.92 ,-1.95]	[-1.95 , -2.26]	[-2.26 ,-∞]
7	[∞ , 5.18]	[5.18 ,4.53]	[4.53,3.00]	[3.00 ,2.98]	[2.98 , 2.32]	[2.32 , 1.58]	[1.58 ,-1.31]	[-1.31 , −∞]

Table 3.2: TTC Bins

The negative Maximum Likelihood Estimator is chosen here because it penalizes prediction errors in accordance to the likelihood with which they appear. Though it does not necessarily minimize the error on a cell level, like a squared difference, the squared difference does not minimize on an overall level. Further, a numerical process for different correlations, ρ, between 0 and 0.3 is tested in 0.001 increments, to find the optimal Z-Factors series as ρ can not be known before a Z-Factor series is calculated. The range for ρ is chosen to reduce the computational power required. Should the optimal ρ be at the upper bound of 0.3, the range is extended. Since the Z-Factor is assumed to have a unit variance, only the possible Z-Factor series with a variance between 0.95 and 1.1 are considered before the Z-Factor with the highest likelihood estimation is chosen. To include this constraint, the limited-memory BFGS optimization is applied. The resulting optimal Z-Factor series is presented in Figure 3.2 and it can be seen that it corresponds with the evolution of ratings shown in Figure 2.1.

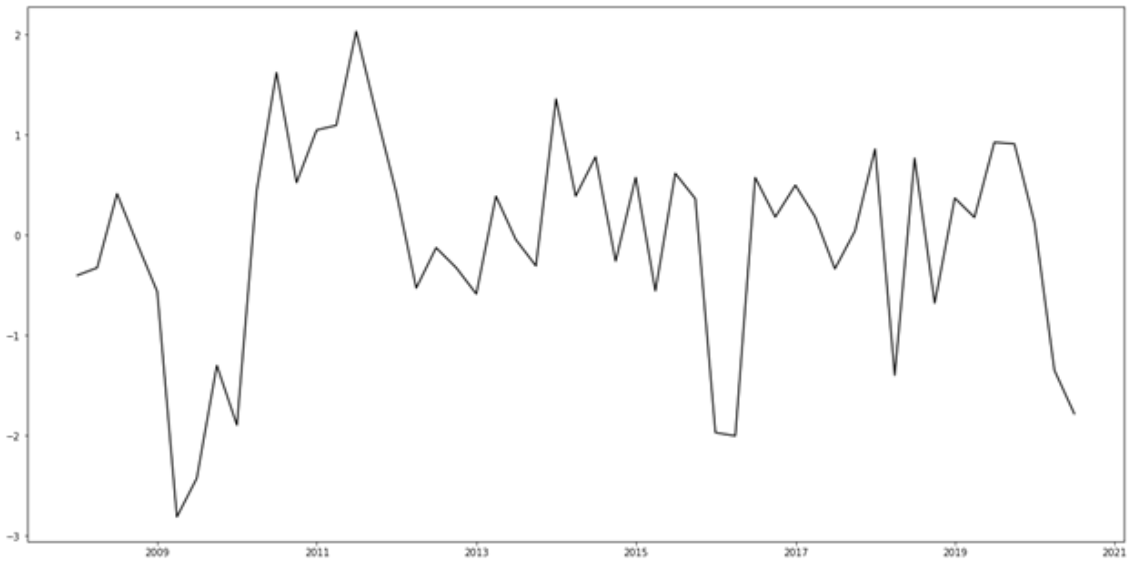


Figure 3.2: Z-Factor time series

3.2 Variable Pre-Selection

The large number of variables considered requires a pre-selection of variables before the final variables can be selected using the respective model. Further, a number of transformations and lags are performed for each variable which increases the number of variables and makes a pre-selection even more important. The selection is done based on criteria such as stationarity and single factor analysis (SFA), similar to Jacobs Jr (2018).

3.2.1 Transformation of Variables

In order to obtain stationary variables, several transformations are applied. For each level variable, the quarterly and yearly difference is computed and for all rate variables, the quarterly and yearly growth is computed. Consequently, for each variable two additional transformations now exist. Due to the fact that some of the macroeconomic variables don't have a contemporaneous relationship with credit risk, as Figlewski et al. (2012) point out, four lags are included for each variable, ranging from one to four quarters. This leads to 14 transformations per variable as seen in Table 3.3.

BBB Spread Lag 3m	*BBB Spread Quarterly Difference*	*BBB Spread Yearly Difference*
BBB Spread Lag 6m	*BBB Spread QD Lag 3m*	*BBB Spread YD 3m*
BBB Spread Lag 9m	*BBB Spread QD Lag 6m*	*BBB Spread YD 6m*
BBB Spread Lag 12m	*BBB Spread QD Lag 9m*	*BBB Spread YD 9m*
	BBB Spread QD Lag 12m	*BBB Spread YD 12m*

Table 3.3: Transformations of the variable *'BBB Spread'*:

3.2.2 Stationarity Selection

Stationarity can be tested using multiple statistical tests. In the scope of this thesis, three tests were considered: The Phillips-Perron (PP) test, the Augmented Dickey-Fuller (ADF) test and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test. The ADF test is part of the unit root tests, designed to test for stationarity. It tests the presence of a unit root, which is the null hypothesis and concludes a series is non-stationary if the null hypothesis cannot

be rejected. To reject the null hypothesis the p-value must be below 0.1. The PP test also belongs to the group of unit root tests. The null hypothesis of the test assumes the presence of a unit root and therefore needs to be rejected for the series to be stationary. The PP test is non-parametric and thus does not require a level of serial correlation as the ADF test does. It is similar to the ADF test though, as it is also based on the Dickey-Fuller test but corrects for autocorrelation and heteroskedasticity. The critical p-value to reject the null hypothesis is also 0.1. The KPSS test differs from the PP test and ADF test because it tests for trend stationarity and not level stationarity. It is based on a linear regression and examines whether a time series is stationary around a mean or a deterministic trend. Similar to the ADF and PP tests, the KPSS test is testing for the presence of a unit root but opposite to the ADF test, the alternative hypothesis represents the presence of a unit root, so that the null hypothesis must not be rejected for stationarity. The critical p-vale is once again 0.1. For each variable transformation, the three stationarity tests are performed. The variables that pass at least two tests are considered as stationary while the other ones are considered non-stationary and discarded.

3.2.3 Single Factor Analysis

For the single factor analysis, only the stationary variables are considered. The goal of the single factor analysis is to determine the direction of the relationship between each explanatory variable and the independent variable, the Z-Factor, and compare it with the expected direction between the two. The expected direction of the relationship is based on common economic understanding. To determine the relationship between a variable and the Z-Factor, a simple linear regression is used. If the coefficient is statistically significant with a p-value below 0.05, its sign is compared with the expected direction of the relationship. Only variables where the sign of the coefficient match the expected direction are further considered. This step ensures that only variables are considered where the relationship can be explained by economic theory.

3.2.4 Selection of Transformation of Variable

At this point, it would be possible to estimate a model using the stationary and economically sound variables. However, there can be multiple transformations of each variable, so in order to reduce the variables considered for the model, further selection is applied. The goal of this selection is to find the best transformation of each variable with multiple transformations remaining. I have chosen three separate measures to select the best transformation.

Spearman Rank Correlation

To select the best transformation for each variable, the Spearman rank correlation with the Z-Factor is calculated. Then the transformation with the highest correlation will be chosen for each variable. Spearman rank correlation is preffered to Pearson correlation as it is less of a linear measure. Strictly linear measures are avoided since the purpose of using an SVR model, is the ability to explore non-linear features. Selecting variables based on a strictly linear relationship therefore might neglect non-linear features.

Random Forest Importance

In order to avoid introducing linearity in the selection, like Spearman correlation, the random forest feature importance is utilized as an alternative for the selection of the best transformation of each variable. A simple random forest algorithm is applied to the feature set and the permutation importance is calculated for each feature. The permutation importance measures the impact a variable has on the R^2, the explanatory power of the model. Then the R^2 of the model is compared with the R^2 of a model where the values of the variable are randomly permuted. A worse performance of the model in terms of R^2, when the variable is permuted, suggest a higher contribution of the variable to the explanatory power of the model. Based on the permutation importance the transformation with the highest score is selected for each variable.

SVR Importance

Similarly, to the random forest importance, the SVR importance is determined by applying an SVR model to the feature set. The permutation importance of each feature is calculated and the transformation with the highest score is selected for each variable.

3.2.5 Multicollinearity

Finally, the pre-selection of variables is further reduced by decreasing multicollinearity between the explanatory variables. The multicollinearity is measured using the variance inflation factor (VIF). The VIF quantifies by how much the variance of the coefficient of a variable is increased due to correlation with the other variables in a multivariate regression. This is calculated by regressing each variable on the other explanatory variables. Then the variance inflation factor for the j^{th} explanatory variable is defined, using the R_i^2 from that regression:

$$VIF_j = \frac{1}{1 - R_i^2} \tag{3.7}$$

With the VIF calculated for all explanatory variables, the variable with the highest VIF is dropped. This process is repeated with the remaining variables until the highest VIF factor is below a threshold of 10. The threshold of 10 is considered as a rule of thumb and was suggested, for example by, Hayden (2005).

3.3 Machine Learning Model

Machine learning algorithms are generally divided into two classes. There are algorithms that solve classification problems and other that are designed for regressions problems. Since the Z-Factor is a continuous and known target variable, the estimation of the Z-Factor is a regression problem. In the scope of this work, I have considered three popular supervised machine learning regression algorithms: Random Forest (RF) Regressions, Support Vector Machine (SVM) Regressions and Neural Network (NN) Regressions. In

the following, I will examine the characteristics and structure of all three algorithms and explain my choice for the SVM Regression. Every machine learning model is built by splitting the available dataset into training and test set. The training dataset is employed for the in-model variable selection, utilizing three different feature importance measures; permutation importance, drop-column importance and Shapley value. To control the learning process, machine learning models possess hyperparameters. The final model will be selected by tuning these hyperparameters through grid search cross validation, which is also done on the training set. Finally, the accuracy of the model is evaluated on the test dataset, using accuracy measures and the explanatory power, R^2.

3.3.1 ML algorithms considered

The RF regression is a tree-based algorithm that uses an ensemble technique to reduce overfitting and the variance. A decision tree is mapping the input features to the target variable by creating decision rules to split the data at different levels. To decide which feature to choose and what condition on the feature to use for a split, the Gini index is used as a measure. For a continuous target variable, the output at the end of each branch of the decision tree is the average of all target variables in the training set that land at the respective end of a branch. This leads to a comprehensible and non-linear model. Such decision trees are sensitive to the data they are trained on and therefore prone to overfit. To prevent overfitting, Breiman (2001) proposed an ensemble method, known as random forests. Random forests utilize the bagging technique, to combine many decision trees in one model. Bagging or bootstrap aggregation creates a number of subsamples with replacement and then trains a decision tree for each subsample. The output of the model for a continuous target variable is then computed as the average of the individual tree outputs. These trees can be structurally very similar with a high correlation of the outputs. Therefore, at each split the random forests randomly limits the pool of features, to select from. This reduces the correlation between the trees in the random forest. Due to its construction, the random forest regressor provides a potentially efficient estimator for

the Z-Factor that can explore non-linearities and is transparent with feature importance measures that can explain the algorithms decisions. However, this construction brings with it, a big drawback. Due to the averaging of the output of the individual trees, it is not possible for random forest algorithms to extrapolate. For the application of the Z-Factor estimation this presents a crucial drawback, as the model will not be able to predict a Z-Factor outside of the past observation. This is an issue that is yet to be resolved. Zhang et al. (2019) provide an approach, where they combine a linear regression with a random forest algorithm to solve the extrapolation issue. They build a linear regression model and use a random forest to explain the non-linearities in the residuals of the regression. Since the variable selection and the majority of the explanation is using the linear model, I have not considered this as a fitting model to explore non-linearities.

Neural network algorithms are often referred to as deep learning algorithms. Their structure is built to resemble the human brain, consisting of nodes that are interconnected like neurons in the brain. These nodes are set up in different layers. The input layer, with a node for each input feature, the hidden layers with an arbitrary number of nodes and finally the output layer, with one node representing the result in case of a regression problem. The deep learning refers to the hidden layers of nodes that can consist of millions of nodes and can produce very powerful models but as the name suggests, the explanation of such a model is very difficult, as it cannot be retraced how the model arrives at its output. The nodes assign weights to each input they receive and when data is feed through the network, the weights are multiplied with the input and added, so that the node gives one number to the next node. This can be suppressed though if the number is below a certain threshold. When the model is trained, these weights and threshold, which are random initially, are learned for each node. This little initial structure leads to neural networks being prone to overfitting and requiring large datasets to be trained on, to ensure that they give consistent results. Additionally, there are 6 hyperparameters that also require a large amount of data to be optimally tuned. Due to the limited data available, neural networks do not present an appropriate algorithm for the Z-Factor estimation.

Support vector machines classify data by creating decision boundaries between ho-

mogenous groups. The algorithm takes the input space and translates it into a feature space of higher dimension in order to perform linear separation. These boundaries are called hyperplanes. In this work, I will focus on the so-called ε-intensive support vector machine, introduced by Vapnik et al. (1997). This algorithm allows for a margin of error when fitting data to theses hyperplanes. This margin is called epsilon and observations that lie in the margin of error are not considered when the error is minimized. Finally, a constant C determines the trade-off between the flatness of the function and the margin of error allowed, by penalizing errors outside of the tolerated margin. In the case of a regression problem, like the one at hand with the Z-Factor estimation, the algorithm resembles a linear regression in a higher dimension as the framework by Basak et al. (2007) describes. The ε-intensive support vector regression (SVR) looks for a function $F(x)$ that allows a margin of error of maximum ε from the target variables using the training data, while trying to keep the function as flat as possible. A linear kernel function looks as follows:

$$F(x) = \langle w, x \rangle + b \tag{3.8}$$

with $w \in \aleph$ and $b \in \Re$, $\langle ... \rangle$ is the dot product. Then the optimization problem can be described as:

$$\min \frac{1}{2}|w^2| \tag{3.9}$$

$$\text{s.t.} \begin{cases} y_i - \langle w, x_i \rangle - b) \leq \varepsilon \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon \end{cases} \tag{3.10}$$

This optimization problem is only feasible though, if all observations lie in the error margin of the approximated function. With observations outside of the margin of error,

the optimization problem becomes the following, with a slack variable, ξ:

$$\min \frac{1}{2}|w^2| + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{3.11}$$

$$\text{s.t.} \begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \tag{3.12}$$

Here the constant C is introduced as a penalization term for observations outside of the allowed margin and therefore represents the trade-off between this error and the flatness of the function f. It allows for the regression to fit a line and boundaries with a worse fit to certain observations but a better fit for the majority of observations. Figure 3.3 visualizes the optimization problem. The ε-intensive loss function that gives this type of SVR its name, consequently looks as follows:

$$|\xi|_\varepsilon = \begin{cases} 0 \quad if \quad |\xi| < \varepsilon \\ |\xi| - \varepsilon \quad \text{otherwise} \end{cases} \tag{3.13}$$

Extending this framework to non-linear kernel functions, dual formulation is applied, as presented by Basak et al. (2007). The nonlinear kernel function can take polynomial form:

$$K\langle X_i, x_j \rangle = (\varepsilon \langle x, x' \rangle + r)^d \tag{3.14}$$

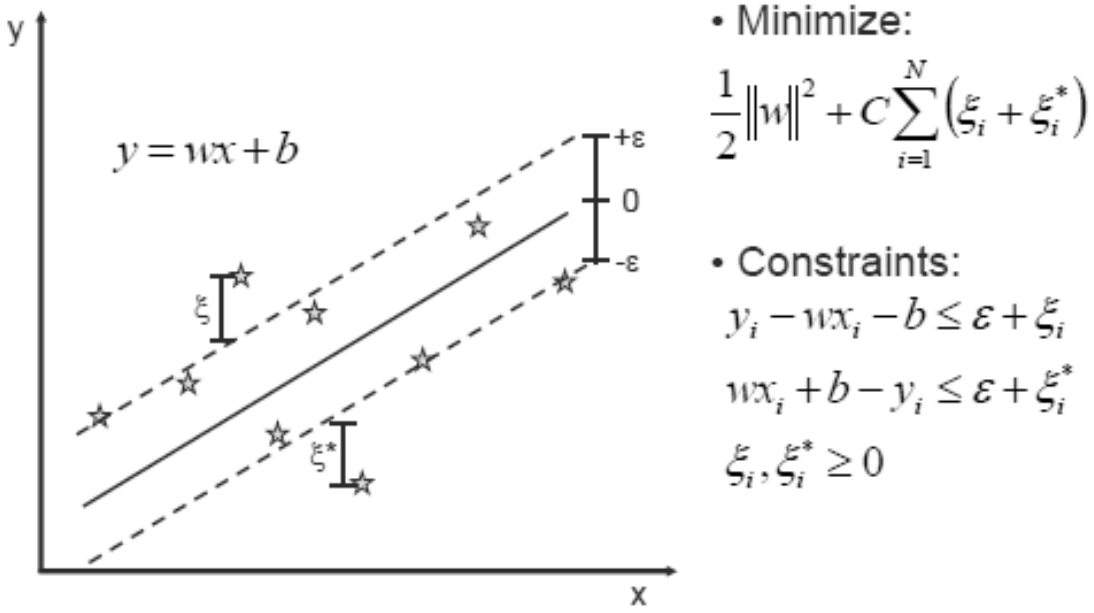
As well as sigmoid form:

$$K\langle X_i, x_j \rangle = (\tanh(\varepsilon \langle x, x' \rangle + r)) \tag{3.15}$$

And finally a gaussian radial basis function (RBF) can be used for the kernel:

$$K\langle X_i, x_j \rangle = exp(-\varepsilon |x - x'|^2) \tag{3.16}$$

A visualizing of the nonlinear kernel can be seen in Figure 3.4. This shows the SVM process of transforming a nonlinear feature space into a higher dimension linear space. Mapping features into a higher dimension can be computationally very expensive but the



Source: `https://www.saedsayad.com/support_vector_machine_reg.htm`

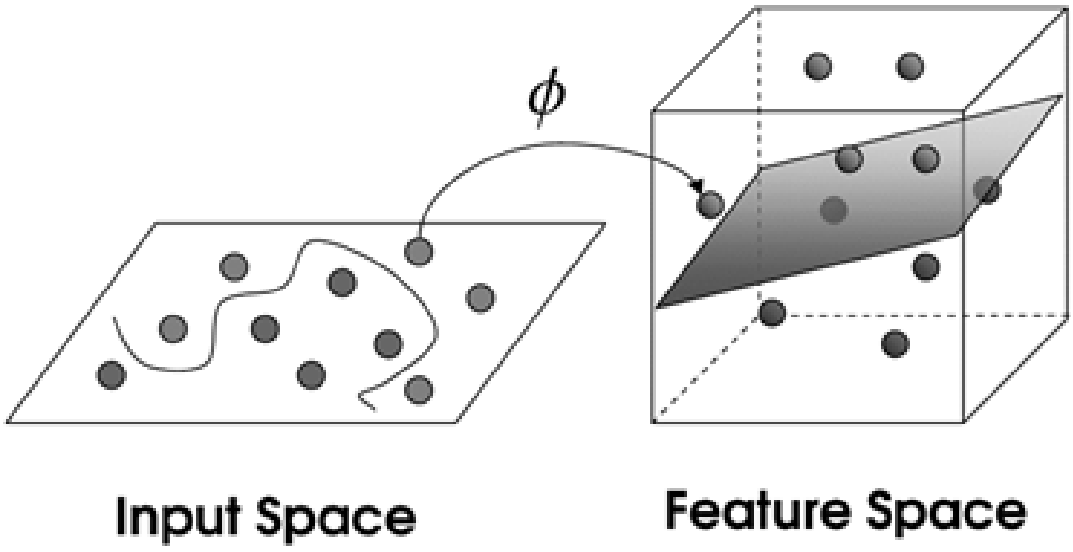
Figure 3.3: Linear SVM

kernel functions makes it possible to directly calculate a non-linear relation between the features.

This kernel trick makes the exploration of non-linearities possible which is an important part of this thesis. SVR models allow forecasting and extrapolation as shown in Guajardo et al. (2006) which are other important characteristics in the scope of the Z-Factor estimation and is not possible, for example with random forests. This is the main reason for my choice of the support vector regression algorithm.

3.3.2 Variable Selection

Before the variable selection is performed, the data is divided into test and training dataset. The training dataset will be used to select the variables of the model, while the test dataset will be held out until the final evaluation of the model candidates. The training dataset represents 70% of the observations and ranges from 2007-2016. The test dataset contains the remaining 30% of observations from 2016-2020. The initial hyperparameters of the



Source: `https://www.jeremyjordan.me/support-vector-machines/`

Figure 3.4: Non-linear SVM

SVM model, used to perform the variable selection, are set according to the suggestion by Cherkassky and Ma (2004). Following their approach, the selection of C, the penalization of errors, depends solely on the target variable:

$$C = |\bar{y}| + |3\sigma_y| \tag{3.17}$$

The selection of epsilon depends on the number of training samples as well as the noise of the input data, σ:

$$\varepsilon = \frac{3\sigma}{\sqrt{\ln(n)/n}} \tag{3.18}$$

The noise is estimated with the residuals of a simple linear regression of the target variable on the explanatory variables. The noise is then defined as the sum of squared residuals, adjusted for sample size and degrees of freedom:

$$\hat{\sigma}^2 = \frac{n}{n-d}\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.19}$$

To select the initial kernel, an experimental approach is chosen, as suggested by Guajardo et al. (2006). The four possible kernels, linear, polynomial, sigmoid and RBF are

tested and the kernel with the lowest mean absolute error and mean squared error is selected. The gamma for the RBF and sigmoid kernels is set as 1 over the number of variables:

$$\gamma = \frac{1}{k} \tag{3.20}$$

Machine learning algorithms are often considered black boxes and it is indeed much more challenging to attribute the outcomes of a machine learning model to the input variables than it is, for example, for a linear regression model. However, there are certain methods that can be implemented for machine learning models that attribute an importance to the input variables.

Permutation Importance

The permutation importance measures the contribution of a variable by randomly permuting a variable's values. The model is then estimated with the permuted variable and the R^2 of the model calculated. Next, this R^2 is subtracted from the R^2 of the model without any permutation. Consequently, a large difference between the R^2s indicates that the model's explanatory power is reduced when the variable is permuted, while a very low or negative difference suggests the model's explanatory power does not or barely depends on the variable. The random permutation is repeated 100 times, independently for each variable to obtain a consistent score. Then, the variables can be ranked by their average importance of the 100 repetitions.

Drop-column Importance

The drop-column importance measure is computed similarly to the permutation importance. Instead of permuting a variable's values, the drop-column importance is calculated, estimating the model without the variable in question. Then, the difference between the R^2 in this model and the original model, including the variable, can be calculated. Again, a higher positive difference indicates the variable provides an important contribution to

the model's explanatory power. Each variable is left out of the model once to determine its drop-column importance, according to which the variables can then be ranked.

Shapley Value

The Shapley value has its origin in game theory where each variable represents a player and the prediction represents the payout. The Shapley value aims to quantify how much each variable contributed to the prediction compared to the average prediction. This can be done based on the method introduced by Shapley (1953) who assigned a payout to players according to their contribution. The contribution of a variable according to this method is calculated as the average absolute difference in the target variable for each of the variable's values, with all other variables' values constant at one possible combination. This is repeated for all possible combination of holding the other variables constant and the average over all contributions is taken. This process is performed for all variables which allows to rank the contribution of each variable with the Shapley value. The Shapley value, $\phi_j(val)$, is mathematically defined as follows:

$$\phi_j(val) = \sum_{S \subseteq \{x_1,\ldots,x_p\} \ \{x_j\}} \frac{|S|(p-|s|-1)!}{p!}(val(S \cup \{x_j) - val(S)) \tag{3.21}$$

where S is one possible combination of variables, x are the variable values corresponding to the p variables selected and $val_x(S)$ is defined as:

$$val_x(S) = \int \hat{f}(x_1,\ldots,X_p) dP_{x \notin S} - E_X(\hat{f}(X)) \tag{3.22}$$

With three different importance measures for each variable, the 3-4 variables that are ranked highest across these measures are selected as the input variables for the model. The reason for the selection of 3-4 variables is on the one hand to maintain the explanatory power of the model and on the other hand the limited observations available with 36 data points in the training set. This allows for roughly 10 observations per variable with 3-4 variables selected which is the general rule of thumb for multivariate regression problems based on Harrell (2017). Since the variable ranking is not always consistent across importance measures, different variable sets are tested on their performance.

Hyperparameter Tuning with Cross Validation

Once the features for the model are selected, setting the optimal hyperparameters is the last step in the model building process and optimization. I have deployed grid search, an exhaustive method, to find the optimal hyperparameters. This method can be computationally expensive but not in the the case of this model. Since the training dataset is relatively small with 3-4 variables and 36 observations, and there are only 2 or 3 hyperparameters to tune, depending on the kernel: the kernel, C, ε and γ for RBF and sigmoid kernels or the degree for the polynomial kernel. The grid search is applied over the 4 kernels with a range of possibilities for each parameter.

$$C = [1,2,3,4,5,6,7,10,15,20,100,1000] \tag{3.23}$$

$$\varepsilon = [0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8] \tag{3.24}$$

$$\gamma = [1e^{-5},1e^{-4},0.001,0.01,0.1,0.2,0.5,0.6,0.9] \tag{3.25}$$

$$degree = [2,3,4,5,6,7] \tag{3.26}$$

Since the number of observations is small, cross validation is implemented. Cross validation uses the available data more efficiently by splitting the data into training and test set multiple times. Since the data is time sensitive, it is important these splits don't happen randomly, as it is the case for the popular k-fold cross validation. To respect the time component and train the model to make forecasts, the initial training set of 36 observations is split into an initial set of the first 8 observations to predict the next 7 observations. Then these 7 observations are included in the training set to predict the next 7 and so on. This can be repeated four times with the training set size of 36.

To decide between models, a scoring parameter has to be defined. I have used the means squared error (MSE), mean absolute error (MAE) and adjusted R^2 as such a parameter. For each parameter setting, there are four scores, one for each cross validation split. The average of these is taken to obtain one score per parameter setting. Then the model with the lowest score is selected in the case of the mean squared error and the mean absolute error and the model with the highest score is selected in the case of R^2.

3.3.3 Model Evaluation

To evaluate the final model, the out-of-sample test set that has been put aside so far, is utilized. Three measures are used to evaluate the accuracy of the model on the out-of-sample data. The MSE takes the mean of the squared difference between the prediction and the actual observation of the target variable, the Z-Factor. Thus, the MSE puts more weight on large deviations. The MAE gives the same weight to all errors by taking the mean of the absolute difference between the prediction and the actual observation. Finally, the adjusted R^2 is used to measure the percentage of variation explained by the features, while controlling for the number of features.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{3.27}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{3.28}$$

$$Adjusted\ R^2 = 1 - [\frac{(1-R^2)(n-1)}{n-k-1}] \tag{3.29}$$

3.4 Linear Model

The linear model will serve as a benchmark, mainly to evaluate the ability of the SVR model to explore non-linearities in the variable selection process as well as the performance of the different model types. The benchmark model is applied to the same set of pre-selected variables as the machine learning model. This means only stationary variables that passed the single factor analysis are considered. Further, the VIF filtering is applied to reduce multicollinearity. This results in the first variable set. As for the machine learning model, another set of variables is constructed by selecting one transformation per variable. Similar to the Spearman correlation selection previously, the transformations are selected according to the highest Pearson correlation. The Pearson correlation is used here as the criterion because this is a linear model, so a linear pre-selection is appropriate. To reduce multicollinearity in this variable set as well, the VIF filtering is also applied here.

Before the model varaiable selection is applied, the explanatory variables and the independent variable are standardized, to allow for coefficients to be interpretable. The mean and variance of each variable is calculated and then the variables are standardized by subtracting each observation by the mean and dividing by the standard deviation. This leads to the data being centered around zero with a variance of 1, so it has the properties of a standard normal distribution.

3.4.1 Lasso Variable Selection

Despite the pre-selection, the variable sets used for the linear model are still too large to run a significant linear regression and further variable selection is necessary. The machine learning model utilizes three different importance measure to do this selection, as described above. For the linear model, I will make use of the lasso regression, first introduced by Tibshirani (1996). Lasso stands for least absolute shrinkage and selection operator and is a linear regression with shrinkage. The use of shrinkage results in the coefficients of unimportant variables shrinking to zero. It minimizes the sum of squared residuals, like OLS, but adds a penalty term to the loss function, as seen in the equation below.

$$\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3.30}$$

This technique is also known as L1 regularization where the penalization consists of the sum of absolute coefficients and a shrinkage parameter, lambda. This optimization leads to coefficients with the value of zero. Clearly the lasso regression equals an OLS regression when lambda is zero. Further, a higher lambda causes a higher penalization of coefficients and therefore, the optimization will result in more coefficients with a value of zero. An increase in lambda and accordingly less non-zero coefficients increases the bias but decreases the variance.

Consequently, the lasso regression strongly depends on the shrinkage parameter, lambda. Therefore, hyperparameter tuning is used to find the optimal lambda. Similar to the hyperparameter tuning for the machine learning model, a grid search over a range of possible

lambdas is utilized. The considered range for lambda starts at 0 and takes steps of 0.1 until 10. If the optimal lambda is below 1, a more detailed range from 0 to 1 in intervals of 0.01 is applied, to select the optimal lambda. To find the optimal lambda, a 4-fold time series cross validation like in the machine learning model is used with the MSE and MAE as scorers. If the selected lambdas are different between the MSE and MAE cross validation, the average between the two is taken.

Now that the optimal lambda is known, the lasso regression can be applied to the variable set to obtain the non-zero coefficients. If there are non-significant variables in the regression these are removed by step-wise backward selection, until all coefficients are statistically significant. This represents an additional model alternative.

3.4.2 OLS Model Evaluation

The variables selected with the lasso regression are then used to perform an OLS linear regression. The OLS regression is fit to the training dataset and the model is evaluated on the out of sample test set. The split between training and test set is the same as for the machine learning model. Further, the same three evaluation measures are used in MSE, MAE and adjusted R^2

3.4.3 Gauss Markov Assumptions

Finally, the Gauss Markov assumptions of the model are tested to ensure the OLS regression is the best linear unbiased estimator (BLUE). First, the Z-Factor and the predicted Z-Factor are plotted to establish, that the relationship between the explanatory variables and the independent variable is indeed linear. Next, the absence of multicollinearity is tested using the VIF as described above. If the VIF for all explanatory variables is below 5, the absence of multicollinearity can be assumed. The homoscedasticity of the residuals is tested using the Breusch-Pagan test and the Goldfeld-Quandt test. The Breusch-Pagan test is defined as and approximately follows a chi-square distribution. The null hypothesis of homogeneity is accepted for p-values larger than 0.05. The Goldfeld-Quandt test

compares the sum of squared residuals of two subsets regression on the data and performs an F-Test to determine whether the two differ significantly. Consequently, the null hypothesis represents homogeneity and is accepted for p-values larger than 0.05. To test the normality of the residuals, the Jarque-Bera test is utilized which is based on the kurtosis and skewness. The null hypothesis assumes normality with the kurtosis and skewness equal to 0 and can be accepted with a p-value above 0.05. Finally, the Durbin-Watson test is used to test for autocorrelation in the error term. The presence of autocorrelation would not lead to a bias in the estimation but influences the standard error of the estimator and it would therefore not be BLUE, as it is not the lowest variance estimator anymore. A test statistic between 1.5 and 2.5 indicates there is no autocorrelation in the error term. To assure the relationship between the explanatory variables and the target variable is linear, the predicted and actual observations are plotted.

3.5 Covariate Shift Adaptation

In machine learning it is commonly assumed that the training and test datasets follow the same probability distributions. However, in many applications this is not the case and it is therefore important to test this assumption. If there is a drift in the probability distribution between training and test set, the covariate shift adaptation provides a method to correct this drift. The drift in the data can be detected with a classification mechanism that is described below. Once a drift is detected, it is important to find importance weights to adjust for the different distributions when fitting the model. Sugiyama and Kawanabe (2012) propose the use of the density ratio between the test and training set as an importance estimation. To estimate the density ratio, I estimate the individual densities for test and train set using a Gaussian kernel density estimation. Then the estimates for the test set are divided by the training set estimates to obtain the density ratio. This density ratio is then used as the sample weights in the model cross validation.

3.5.1 Data Drift Detection

To determine drift between the training and test sample, a new target variable is created. This variable is set at 0 for the training set and 1 for the test set. The estimation of this variable presents a classification problem and the same independent variables are used for the estimation. If these variables classify the target variable with a high accuracy, it is an indication for the presence of a drift in the data. I use a simple random forest classifier to determine the accuracy.

3.5.2 Density Ratio Estimation

To estimate the density ratio or importance weight, the density for the training set and for the test set are calculated separately. To do so, the Gaussian kernel density estimation is used. The gaussian kernel density estimator has a hyperparameter, the bandwidth, that requires to be optimized. Leave-one-out cross validation on the respective dataset is used to optimize the bandwidth. With the optimal bandwidth known, the density function of the test and training set can be calculated respectively. Finally, the density ratio is computed as the test density function divided by the training density function, both applied to the training set.

3.5.3 Sample Weight Implementation

Once the density ratio is known, the new model with covariate shift adaptation can be constructed. As described in Sugiyama and Kawanabe (2012), the covariate shift can be applied to an e-intensive support vector regression. They call this method adaptive importance-weighted support vector regression (AIWSVR).

$$\hat{\theta}_\gamma = \underset{\theta}{argmin} \left[\frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \left(\frac{p_{te}(x_i^{tr})}{p_{tr}(x_i^{tr})} \right)^\gamma |\hat{f}(x_i^{tr};\theta) - y_i^{tr}|_\varepsilon \right] \tag{3.31}$$

Equation 3.31 shows the corresponding loss function, where the term $\frac{p_{te}(x_i^{tr})}{p_{tr}(x_i^{tr})}$ represents the density ratio. The power of the density ratio, γ, is assumed to be 1, to fully incorporate

the different training and test distribution. The adjusted loss function is now implemented in the cross validation, previously described to select the optimal model. This is now called an importance weighted cross validation and is almost unbiased as Sugiyama and Kawanabe (2012) show.

Chapter 4

Results

In this chapter, I will present the results of the application of the models, introduced in the previous chapter, to the data, described in chapter 2. I will begin with the results of the variable pre-selection process. Then, I will show the results of the different models, starting with the SVR model. I will display the model performance of the linear benchmark model and the covariate shift model. Finally, I will briefly go over the results of the model with monthly data.

4.1 Variable Pre-Selection

The first attained results are the macroeconomic variable sets obtained from the different variable pre-selection methods. The first results will display the difference in these sets before and after the VIF filtering is applied. For the variables, listed in Table 4.1, no stationary transformation that passed the single factor analysis is found.

When the Spearman correlation pre-selection is applied, the VIF filter removes three variables: Initial Claims QGrowth Lag9m, DJ Total SM Index YGrowth Lag3m, Lending TS LM Corps Lag9m. For the random forest pre-selection, the variables Nominal GDP Growth YD Lag6m, DJ Total SM Index YGrowth Lag3m are filtered out by the VIF. With the SVR pre-selection method, the variables DJ Total SM Index YGrowth Lag3m, Total Assets QGrowth Lag3m, Real GDP Growth Lag3m, Liq. Swaps w/ CBs Lag3m, Loans

Real Disposable Income Growth
Unemployment Rate
Continuous Jobless Claims
House Price Index
WTI Price
M2
Repos
U.S. Treasuries
MBS
DSR (PNFS, Households & NPISH)
DSR (PNFS)
US Dollar Index
10 year - 3 month Treasury Spread
5 year - 3 month Treasury Spread

Table 4.1: Variables without stationary Transformation that passes the SFA

Lag3m are dropped by the VIF filtering. The lists of selected variable sets can be found in the Appendix in Tables A1 - A3.

4.2 SVR Model

In this section, the results for the machine learning model using quarterly data are presented. This includes the model performance comparison between the different variable sets regarding the impact of the pre-selection method and the VIF filter. Further, the kernel and hyperparameter selection are shown as well as the model performance in and out-of-sample. Finally, the variable importance will show the relative impact each of the selected variables have on the estimation. The initial hyperparameters for the model per variable set are reported in Table 4.2. The penalization parameter, C, is the same for all sets, as it only depends on the target variable. The gamma changes according to the number of variables in the respective set. For all variable sets, the RBF kernel is selected as it provides the lowest MSE and MAE among the four possible kernels. The epsilon varies between all variables sets.

Variable Set	Kernel	C	ε	Gamma
Spearman Correlation Pre-VIF Filter	RBF	3.4	0.19	1/22
Spearman Correlation Post-VIF Filter	RBF	3.4	0.22	1/19
RF Importance Pre-VIF Filter	RBF	3.4	0.13	1/22
RF Importance Post-VIF Filter	RBF	3.4	0.24	1/20
SVR Importance Pre-VIF Filter	RBF	3.4	0.11	1/22
SVR Importance Post VIF Filter	RBF	3.4	0.14	1/17

Table 4.2: Pre-Selection Hyperparameters

4.2.1 VIF Filter performance Comparison

To demonstrate the impact, the VIF filtering has on the variable selection, Table A4 in the Appendix shows the importance ranking for the RF pre-selection method before and Table A5 after the VIF filter. It can be observed that the variable ranking is more consistent across the importance measures after the VIF filter is applied. While the variables, Debt-Service-Ratio, BBB Spread and Profitability are consistently ranked highest before the VIF Filter, the variable selection changes to S&P Industrial, BBB Spread and Export Volume. Especially, the S&P Industrial drop-column importance rank rises substantially, as well as the Shapley values of the Export Volume and BBB Spread. The increased drop-column importance of the S&P Industrial can be explained by the removal of the highly correlated Dow Jones Total Stock Market Index by the VIF filter.

To evaluate the impact of the VIF filter on the performance, Tables A6-A8 in the Appendix show the best performing model before and after the filter for each pre-selection method. For the Spearman correlation pre-selection, the variable selection does not change even though 3 variables are removed by the VIF filter. Consequently, the model performance does not differ. The RF pre-selection model performance does not improve with the variables selected in the VIF filtered set while the SVR pre-selection model performance is better with the variables selected from the VIF filtered set.

4.2.2 Pre-selection performance comparison

The best performing models for each of the three pre-selection methods are shown in Table 4.3 and display some similarities in their selected variables with the BBB Spread selected by each method. Further, the Debt-Service-Ratio, Export Volume and Profitability are selected twice. The optimal hyperparameters for the three cross validation scorers, MSE, MAE and adjusted R^2 are shown in the second row. Depending on the variable selection, the scorers select the same model or slightly different ones. The variables of the SVR importance pre-selection return the same model candidate for all three scorers, whereas the other two variable selection lead to two candidates each. For the Spearman correlation selected variables, model candidate 2, which has a different kernel function, maximizes the adjusted R^2, while model candidate 1 minimizes the MSE and MAE. In case of the RF importance selected variables, the MAE is minimized by a model with a slightly different epsilon. The respective cross-validation score can be found in rows 3-5 of Table 4.3. It can be observed that the kernel functions are linear as in the case of the SVR importance pre-selection, or close to linear, since a sigmoid or RBF with a very low gamma is almost linear. The margin of error, ε, is quite large in case of the RF and SVR importance pre-selection, while the penalization parameter is small. For the Spearman correlation pre-selection, the margin of error and penalization parameter are both small. The rows 6-7 present the out-of-sample performance of each of the model candidates. It can be seen that the RF importance pre-selection model candidates perform significantly worse than the other two models with respect to all measures. The SVR importance pre-selection model performs better than the two Spearman correlation pre-selection model candidates in terms of MSE, MAE and adjusted R^2. The SVR importance pre-selection is also the only one where the same model candidate is chosen by all three cross-validation scorers. For the Spearman correlation pre-selection, the model selected by the MSE and MAE performs better than the one selected by adjusted R^2 based on all measures. The best performing model, from the SVR pre-selection, displays a lower MSE than MAE, suggesting that there are no large errors due to outliers. By squaring the errors the MSE

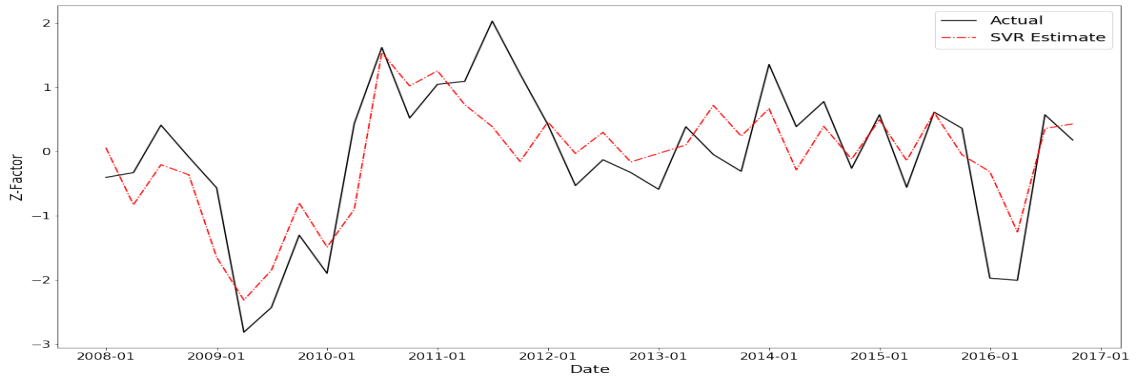


Figure 4.1: In-Sample Performance of best SVR Model

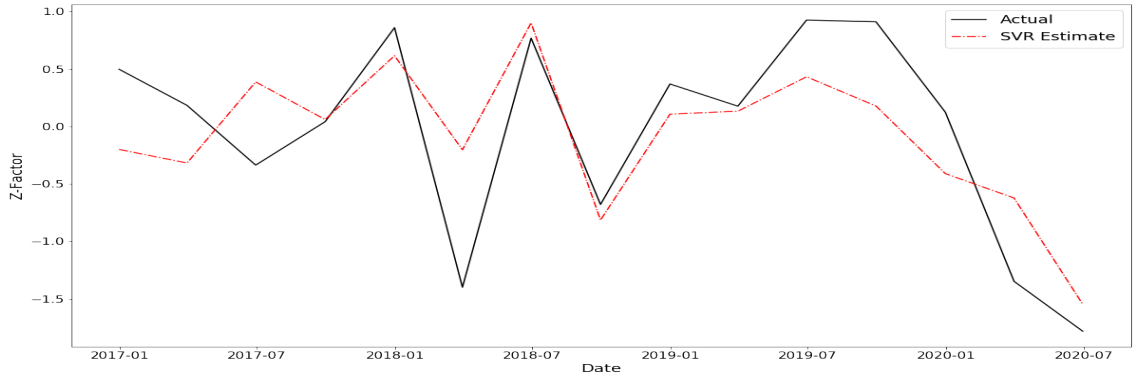


Figure 4.2: Out-of-Sample Performance of best SVR Model

puts more weight on large errors but their presence can be ruled out here as the MSE is low compared to the MAE. This is also confirmed in the Figure 4.2 where the largest deviation is in Q1 2018. To put these errors into context, a look at how they relate to the Z-Factor values is helpful. The Z-Factor for the whole observation period ranges from -3 to 1.5 with an average close to 0. The standard deviation is 1.05 and therefore the MAE of 0.45 represents a deviation of the prediction from the actual value of a little less than half a standard deviation on average.

The fit of the best performing model, the SVR importance pre-selection model, is visualized for the in and out-of-sample period in Figures 4.1 and 4.2. The largest discrepancies between estimation and actual Z-Factor occur in Q3 2011 and Q1 2016 during the energy crisis in the in-sample and in Q1 2018 out-of-sample. The in- and out-of-sample comparison between the Z-Factor and its estimate for the other two best performing pre-selection

Transformation Pre-Selection	Spearman Correlation	RF Importance	SVR Importance
Variables	US export goods vol QGrowth,	NIPA Profitability QGrowth Lag12m,	NIPA Profitability non financial QGrowth Lag12m,
	DSR Corps non financial QGrowth Lag12m,	DSR Corps non financial QGrowth Lag12m,	Commercial RE PI QGrowth,
	BBB Spread QD Lag12m	BBB Spread QD Lag12m,	BBB Spread QD Lag12m,
		Loans QGrowth Lag12m	US export goods vol QGrowth
Hyperparameters			
Candidate Model 1	Kernel=sigmoid, C=3, ε=0.1, γ=0.001	Kernel=sigmoid, C=1, ε=0.6, γ=0.0001	Kernel=linear, C=1, ε=0.5
Candidate Model 2	Kernel=rbf, C=2, ε=0.1, γ=0.001	Kernel=sigmoid, C=1, ε=0.8, γ=0.0001	
CV-MSE	1.0340	3.9498	0.4579
CV-MAE	0.7258	0.8581	0.5382
CV-adj. R^2	0.174	-2.1501	0.6348
OOS Performance - Candidate Model 1			
MSE	0.4	0.71	0.3
MAE	0.5	0.71	0.45
adj. R^2	0.49	0	0.62
OOS Performance - Candidate Model 2			
MSE	0.43	0.74	
MAE	0.50	0.93	
adj. R^2	0.46	0	

Table 4.3: Model Results for each Pre-Selection Method

methods can be found in A4 - A9 in the Appendix Figures. For the Spearman correlation pre-selection model, the estimation shows little variation and therefore presents more an average of the actual Z-Factor. The bad performance of the RF importance pre-selection model can be clearly observed in the graphical representation of the Z-Factor and its estimate. The estimation shows large swings, especially during the time of the Great Financial Crisis, between 2008 and 2010. This can be explained by the variable Loans as seen in Figure A8. This variable represents the loan facility of the federal reserve and had extreme variations during the Great Financial Crisis. The estimates do not fit the actual Z-Factor well, as the performance measures in Table 4.3 suggest

4.2.3 Cross-Validation

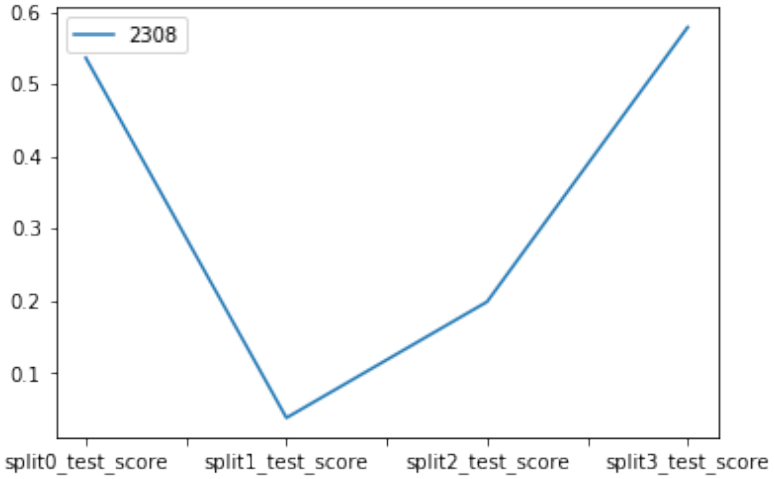
The cross-validation shows a similar pattern for all model candidates. A graphical presentation of the cross-validation scorer over the splits for the SVR-importance pre-selection is shown in Figure 4.3. The graphical presentation for the other models can be found in Figures A1 - A2 in the Appendix. In case of the MSE and MAE, the error is reduced with every split until the last split. In contrast, the adjusted R^2 increases with every split after the first one. This suggests the cross validation is working well as the model is improving with more data added. The lack of improvement in the last split can be explained with a visualization of the actual Z-Factor and its estimation as seen in Figure 4.1. The last cross-validation split contains the 2016 energy crisis and this outlier is not estimated well by the selected macroeconomic variables.

4.2.4 Variable Importance

The variable importance measures in Table 4.4 show that the Commercial Real Estate Price Index has the biggest impact on the estimation followed by the Export Volume, the BBB Spread and the Profitability for the SVR importance pre-selection model. All the variables have a significant impact on the model with the high importance measures, indicating most of the variation in the estimation can be explained by the variables. The rank-

(a) MSE CV of best SVR Model

(b) MAE CV of best SVR Model



(c) R^2 CV of best SVR Model

Figure 4.3: Cross-Validation Results of best SVR Model

ing is also consistent across the measures except for the Shapley value but here the differences between variables are minor anyways. A graphical representation of the macroeconomic variables and the Z-Factor can be found in Figures 4.4 and 4.5. All variables are standardized, to make the macroeconomic variables comparable with the Z-Factor. The sign of the effect for each variable can already be inferred by observing the co-movement between the variables and the Z-Factor in these figures. Additionally, the Shapley value also offers insight into the relationship through dependence plots. These are shown in the Appendix in Figure A20 and confirm that the relationship between the Z-Factor and the explanatory variables is positive except for the BBB Spread which is in line with the economic understanding. These figures can explain the main discrepancies observed between the estimation and the actual Z-Factor. As previously mentioned there is a lack of fit for Q3 2011, with the estimate pointing lower that the actual Z-Factor. This is mainly caused by the low Commercial Real Estate PI which has the largest impact according to the importance measures. The Profitability also reduces the Z-Factor estimation but as seen in Table 4.4, it has a much smaller effect on the estimation. For the other major discrepancy between estimated and actual Z-Factor in the in-sample can be found in Q1 2016 where the variables all point in the right direction but their magnitude is not large enough which results in the estimate being less negative than the actual Z-Factor. In the out-of-sample the main divergence between estimated and actual Z-Factors can be found in Q1 2018. Here, solely the Commercial Real Estate PI displays the appropriate move for an accurate estimate. Even though it has the highest influence on the estimation, it is not enough to offset the too small magnitude in the Export Volume move down. Further, the high Profitability and low BBB Spread lead to a higher estimate for the Z-Factor.

The variable importance measure and the graphical presentation containing the macroeconomic variables for the Spearman correlation and RF pre-selection model can be found in the Appendix in Tables A10 & A11 and Figures A6 - A11. For the Spearman correlation pre-selection, it shows that the BBB Spread has the biggest impact on the estimation, followed by the Debt-Service-Ratio and the Export Volume has the lowest impact. However, the relatively low importance for all the variables shows that none have a large

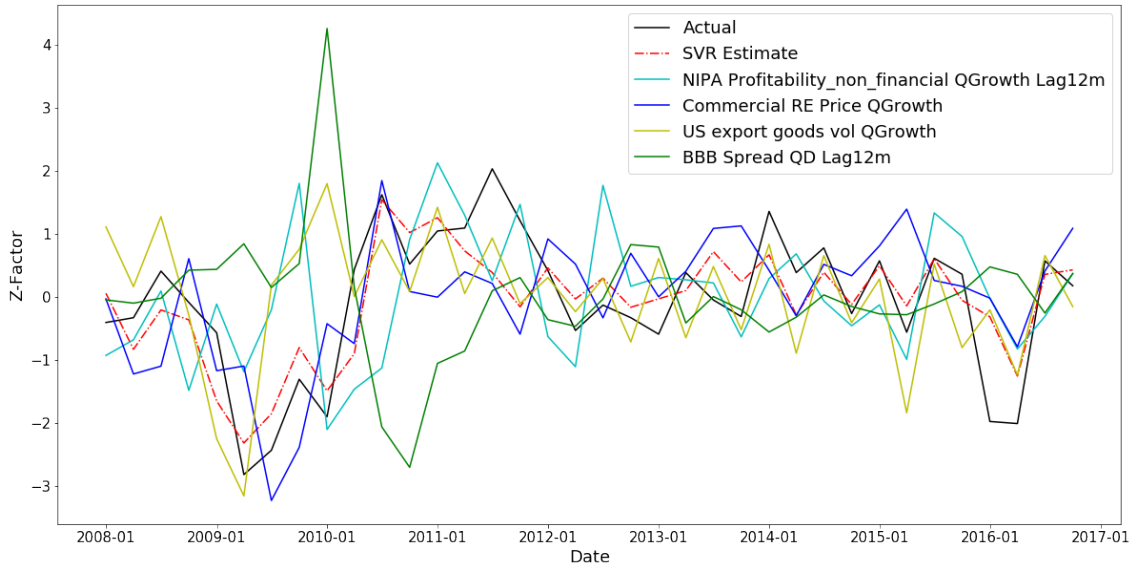


Figure 4.4: In-Sample Performance of best SVR Model with MEVs

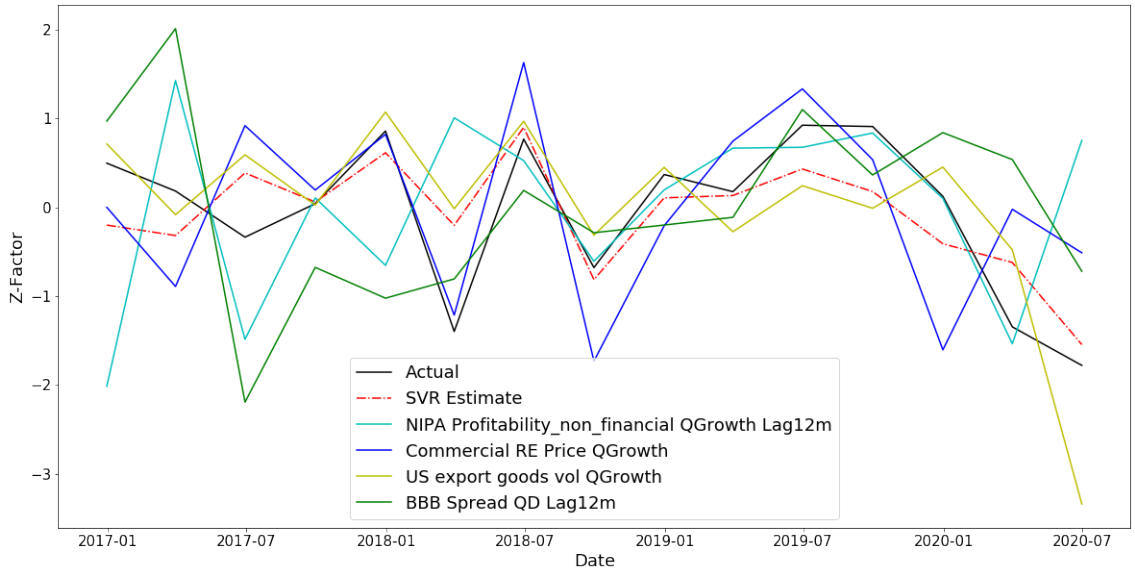


Figure 4.5: Out-of-Sample Performance of best SVR Model with MEVs

Variable	Permutation Importance	Drop-Column Importance	Shapley Value
Commercial RE PI QGrowth	0.43	0.17	0.12
US export goods vol QGrowth	0.21	0.10	0.12
BBB Spread QD Lag12m	0.18	0.09	0.11
NIPA Profitability non financial QGrowth Lag12m	0.13	0.05	0.12

Table 4.4: Variable Importance

impact by themselves. For the RF pre-selection the large impact of the Loan variable is only present at certain point in time, and for the entire time frame, the BBB Spread represents the most important variable as in the Spearman correlation pre-selection model. It is followed by the Profitability, Loans and finally the Debt-Service-Ratio. The importance measure for permutation and drop-column importance are extremely close to zero and indicate none of the variables have a large impact by themselves.

4.3 Linear Benchmark Model

This section will present the results for the linear benchmark model for the two pre-selected variable sets considered in the linear approach; the VIF filtered initial variable set (30 variables) and the Pearson correlation transformation selection and VIF filtered set (20 variables). First, the results of the Lasso cross-validation are presented. Then, I will show the variable selection and model performance for each variable set in detail. Finally, the test results of the Gauss-Markov assumption for the models will be displayed.

4.3.1 Variable Selection

Before the Lasso can be implemented for the variable selection, the optimal lambda for shrinkage must be determined. The results of the cross-validation for each of the two variable sets are shown graphically in Figure A3 in the Appendix. For the VIF filtered variable set, the optimal lambda is 0.02, using the MSE and MAE as the scorer in the cross-validation. In the second variable set the optimal lambda differs between the MSE

and MAE as the scorer. For the MSE it is again 0.02 but for the MAE it is 0.17, so the average, 0.1, is used. Since the lambda is very low for the first variable set, the Lasso variable selection results in a large amount of 22 variables left, as only 8 coefficients are shrunk to 0. Even the further reduction, using the stepwise selection of significant variables, leaves 11 variables in the model. Therefore, I increase the lambda until 4 or 3 variables remain, respectively. This is achieved with lambdas of 0.5 and 0.55. In case of the second variable set, the smaller number of initial variables and higher lambda result in a set of seven variables. After the stepwise selection, only two variables remain as significant: S&P Industrial YGrowth Lag3m and US export goods vol YGrowth Lag3m. As for the previous set, I increase the lambda until only 4 or 3 variables remain to obtain further model candidates. The lambdas of 0.3 and 0.4 achieve this, respectively. The four variable sets selected can be found in the Appendix in Table A9 as well as the four variable sets selected for the VIF filter pre-selection set. It is important to note that the variable sets for the model candidates of the VIF filtered pre-selection include several transformations of the same variable. This displays the importance of the transformation selection which ensure only one transformation per variable is selected. Including multiple transformations of a variable usually adds little informational gain but a lot of noise to the model.

4.3.2 Model Performance

The model candidates with a large number of variables, as one would expect, show a very good fit in-sample with low errors and large adjusted R^2s as seen in the Appendix in Table A9. However, out of sample these models perform poorly with large errors and adjusted R^2s of 0. As mentioned above, using only the VIF filter as a pre-selection leads to several transformations of one variable selected which results in bad model performance for the three models where this is the case. The best performing model in terms of the lowest out-of-sample errors is the one where 3 variables are selected by an increase in lambda and does not include multiple transformations of the same variable. However, the errors are just slightly lower and the adjusted R^2 is 0 as for the other models. For the

second variable set, the model including all non-zero coefficients belongs to the models that display a good in-sample performance and a poor out-of-sample performance due to a large number of variables selected. The other three models show much less discrepancy between in-sample and out-of-sample performance. As for the models resulting from the VIF filtered pre-selection, the models resulting from the second variable set all have an out-of-sample adjusted R^2 of 0 but the errors are significantly lower. According to the different error measures the model with four variables selected, through a lambda of 0.3, performs best as shown in Table 4.5. The error rates at 0.66 and 0.69 for the MSE and MAE, respectively, do not suggest a very good fit. Relating this to the Z-Factor standard deviation, the MAE represents an average deviation between the prediction and the actual Z-Factor of a little more than half the standard deviation. This can also be seen in the graphical presentation of the estimate and the Z-Factor in the Appendix in Figures A12 & A13. The fit is good in the in-sample, especially in the period from 2008 – 2014. Then the estimate becomes more of an average of the actual Z-Factor, as it does not fit the peaks and troughs well which is particularly evident in 2016 with the oil glut crisis. This behavior continues into the out-of-sample and only the deep deterioration in the Z-Factor in Q2 of 2020 is also reflected in the estimate. Since the data is standardized, the coefficients magnitude informs about the weight in the estimation. The S&P Industrial YGrowth Lag3m has by far the biggest influence on the estimation with 0.43. It is followed by the US export goods vol YGrowth Lag3m with 0.21, ISM PMI Services YGrowth Lag12m with 0.18 and Liq. Swaps w/ CBs Lag3m with -0.15. When plotting these variables with the estimated and actual Z-Factor, it becomes evident why the Liq. Swaps w/ CBs Lag3 has the lowest influence as it only differs significantly from 0 in times of stress as during the financial crisis and most recently in 2020. However, in those periods it plays an important role as the estimates in these periods are more accurate. The S&P Industrial YGrowth Lag3m tracks the Z-Factor very well until about 2014 which explains, the discrepancy in accuracy after 2014, as it is the variable with the by far biggest influence on the estimation. The US export goods vol YGrowth Lag3m shows a fairly good fit across the whole sample which justifies the higher coefficient compared to the ISM PMI

Transformation Selection	VIF Filtered	VIF Filtered & significant	Reduction to 4 Variables	Reduction to 3 variables
Variables	S&P Industrial YGrowth Lag3m, Liq. Swaps w/ CBs Lag3m, ...	S&P Industrial YGrowth Lag3m, US export goods vol YGrowth Lag3m	S&P Industrial YGrowth Lag3m, US export goods vol YGrowth Lag3m, Liq. Swaps w/ CBs Lag3m, ISM PMI Services YGrowth Lag12m	S&P Industrial YGrowth Lag3m, US export goods vol YGrowth Lag3m, Liq. Swaps w/ CBs Lag3m
Lambda	0.1	0.1	0.3	0.4
IS-MSE	0.3277	0.4243	0.3669	0.3999
IS-MAE	0.4341	0.5318	0.4812	0.5188
IS-adj.R^2	0.6386	0.5632	0.6345	0.6141
OOS-MSE	0.6821	0.7542	0.6619	0.7155
OOS-MAE	0.7207	0.7395	0.6925	0.7156
OOS- adj.R^2	0	0	0	0

Table 4.5: Linear Model Performance

Services YGrowth Lag12 which fits to a lesser degree specifically in 2016.

4.3.3 Gauss-Markov Assumptions

The results of the Gauss-Markov assumption tests for all the linear model candidates can be found in the Appendix in Table A12. A majority of the models passes these tests and satisfy the assumptions. One exception is the first model from the VIF filter pre-selection set which contains 22 variables. It exhibits multicollinearity and autocorrelation in the residuals and therefore violates 2 of the Gauss-Markov assumptions. The best performing benchmark model passes the Jarque-Bera test for normality by accepting the null hypothesis with a high p-value of 0.95. The 4 VIFs are all well below 5 and the Durbin-Watson test statistic of 2.16 suggests no autocorrelation in the error term. The

homogeneity assumption for the error terms is satisfied at a 5% significance level with a p-value for the Breush-Pagan test of 0.15 and a p-value for the Goldfeld-Quandt test of 0.08. Therefore, all Gauss-Markov assumptions are fulfilled.

4.4 Covariate Shift Adaptation

In this section, I will present the results of the covariate shift implementation for the best performing model of the three pre-selection methods. First, I will present the results for the detection of a drift in the data that would justify a covariate shift adaptation. Then, I will present the results of the covariate shift adaptation to show the effect it has on the model performance and the variable importance.

4.4.1 Data Drift Determination

The results of the classification problem between test and training set to determine whether there is drift in the data is measured by the classification accuracy. The in-sample accuracy is 100% for all three models and the out-of-sample accuracy is very high as well with 92% for the Spearman correlation pre-selection, 94% for the RF importance pre-selection and 90% for the SVR importance pre-selection. This clearly indicates a drift in the data between the training and test set and therefore a covariate shift adaptation is appropriate. This requires the determination of sample weights which are calculated using the Gaussian kernel density estimation as described above. The results of the weights can be found in Figure A21 in the Appendix. It is interesting to note that the observations of the Great Financial Crisis get small weights attributed while the observations during the Oil Crisis get higher weights.

4.4.2 Covariate Shift Models' Performance

In Table 4.6 the results of the covariate shift adaptation for the three models are summarized. The variable selection is not impacted by this adaptation, but it can be observed

that the hyperparameters change. Compared to the previously selected models, the allowed margin of error is smaller for the RF and SVR pre-selection with a higher penalization term, C. The penalization term is also larger than previously for the Spearman pre-selection model but the margin of error increases from 0.1 to 0.2. The functions for all models are again linear or close to linear, as again sigmoid functions with a very low gamma parameter are selected. The RF importance pre-selection model candidates 2&3 with an RBF function are less linear with a gamma of 0.2. However, these models do not perform well once again even though there is an improvement, they are overfitting with the MAE more than doubling from in-sample to out-of-sample for the model candidate 2. The adjusted R^2 for model candidate 3 is drastically dropping from in-sample to out-of-sample also suggesting overfitting. The Spearman correlation pre-selection model does not show any improvement with the covariate shift implementation. The previously best performing model candidates, the SVR importance pre-selection, remains the best performing model and shows further improvement with the covariate shift implementation. The best model, candidate 2, improves the previously best performing model by reducing the MSE by 0.03 and the MAE by 0.1 and increasing the adjusted are square by 0.03. Recalling that the standard deviation of the Z-Factor is 1.05, the MAE of 0.35 presents roughly an average deviation of the prediction from the actual value of as little as a third of a standard deviation. This further improved fit can, for example, be observed at Q2 2020 as shown in the Appendix in Figure A17.

4.4.3 Variable Importance

The variable importance measures allow a more detailed look on how the explanatory variable contribute to the estimation and are shown in Table 4.7. According to the permutation and drop-column importance, the Commercial Real Estate Price Index and Export Volume remain the most important variables, followed by Profitability and then BBB Spread. The BBB Spread provides the least to the estimation according to all measures. Previously this was the Profitability which is now closely ranked to the first two and even

Transformation Pre-Selection	Spearman Correlation	RF Importance	SVR Importance
Variables	US export goods vol QGrowth, DSR Corps non financial QGrowth Lag12m, BBB Spread QD Lag12m	NIPA Profitability QGrowth Lag12m, DSR Corps non financial QGrowth Lag12m, BBB Spread QD Lag12m, Loans QGrowth Lag12m	NIPA Profitability non financial QGrowth Lag12m, Commercial RE PI QGrowth, BBB Spread QD Lag12m, US export goods vol QGrowth
Hyperparameters			
Candidate Model 1	Kernel=sigmoid, C=15, ε=0.2, γ=0.01	Kernel=linear, C=2, ε=0.6	Kernel=linear, C=1, ε=0.3
Candidate Model 2	Kernel=sigmoid, C=20, ε=0.2, γ=0.01	Kernel=rbf, C=100, ε=0.1, γ=0.2	Kernel=sigmoid, C=1000, ε=0.2, γ=0.0001
Candidate Model 3		Kernel=rbf, C=1000, ε=0.5, γ=0.2	
CV-MSE	1.011	3.4247	0.4968
CV-MAE	0.7087	0.26033	0.6166
CV-adj.R^2	0.2028	0.7278	0.5096
OOS Performance - Candidate Model 1			
MSE	0.44	1.98	0.3
MAE	0.55	1.14	0.4
adj. R^2	0.44	0	0.62
OOS Performance - Candidate Model 2			
MSE	0.43	0.71	0.27
MAE	0.55	0.65	0.35
adj. R^2	0.45	0.09	0.65
OOS Performance - Candidate Model 3			
MSE		0.70	
MAE		0.68	
adj. R^2		0.1	

Table 4.6: Performance of Covariate Shift Models

Variable	Permutation Importance	Drop-Column Importance	Shapley Value
Commercial RE PI QGrowth	0.26	0.2	0.2
US export goods vol QGrowth	0.25	0.09	0.32
NIPA Profitability non financial QGrowth Lag12m	0.22	0.06	0.34
BBB Spread QD Lag12m	0.06	0.06	0.06

Table 4.7: Covariate Shift Model Importance Measures

the most important according to the Shapley Value. The Shapley value allows a more detailed look at individual observations such as the last out-of-sample observation, Q2 2020. Figures 4.6 and 4.7 display the contribution of the four explanatory variables to the Z-Factor estimation of -1.55 and -1.76 for the model with the covariate shift adaptation. The variables Export Goods and Commercial Real Estate Price have a negative impact on the estimation, indicated by the blue color, while Profitability and BBB Spread have a positive impact on the estimation, indicated by the red color. When comparing the two model estimations, it can be seen as the Shapley values in the importance measure tables indicate, that the BBB Spread gets relatively less weight in the covariate shift model with the smaller bar, as Figure 4.7 shows. Despite a higher weight given, the Profitability is the main cause for the lower estimation and outweighs the smaller weight given to the Commercial Real Estate Price. The weight of the Export Goods stays roughly the same. This shows how the results of the different models are composed by the explanatory variables.

4.5 Model using Monthly Data

The monthly Z-Factor serious is very volatile which makes the estimation with economic variables difficult. Especially, since many of these variables cannot be used, as they are computed and published on a quarterly basis, the resulting models perform poorly with high errors and low explanatory power of the variance. The monthly approach was originally considered because of the concern that the quarterly observations might not provide

Figure 4.6: Shapley Contributions 2020/Q2 SVR Model

Figure 4.7: Shapley Contributions 2020/Q2 SVR Model with Covariate Shift

enough data points for an accurate estimation. However, this is not the case as the previous results show. It is also uncommon to look at defaults and rating transitions on a monthly basis, since credit cycles last several years, and rating transitions usual do not happen that frequently. For these reasons, I will not further explore this approach.

Chapter 5

Discussion

In this chapter, I will discuss the previously presented results and interpret their meaning in regard to the research question. I will asses the impact of the chosen methodology on the results and explain the economic intuition behind the selected variables. Further, I will compare the model performance between the machine learning model, the linear benchmark model and the covariate shift model.

5.1 SVR Model

In the first section of this chapter, I will focus on the best machine learning model based on the findings in Chapter 4.2. This is the model resulting from the SVR importance pre-selection, presented in column 4 of Table 4.3. It is performing significantly better in all three evaluation measures than the second best candidate from the Spearman correlation pre-selection. Further, all three measures lead to the same model which provides confidence in the hyperparameter specification.

5.1.1 Model Performance Evaluation

VIF Filter

The reason for the application of the VIF filter on the pre-selection sets is the reduction in multicollinearity, before the importance measures are applied because high multicollinearity between variables distorts their drop-column and permutation importance. This is due to these two measures relying on the difference in R^2 when the variable is dropped or permuted. For highly correlated variables this difference and therefore, the importance is smaller than for less correlated variables. Thus, the ranking is more reflective of the impact a variable has individually after the VIF filter application. This has the biggest impact in the SVR pre-selection as evident by the better model performance for the VIF filtered set. Here, 5 variables are removed, indicating multicollinearity present in the initial variable set.

Between pre-selection methods, the SVR was shown to outperform the RF and Spearman correlation methods. This is plausible as the final algorithm is an SVR algorithm and using it in the pre-selection is a coherent approach. The RF algorithm represents a quiet different mechanism as it is based on a decision tree and not a kernel function which leads to a disconnect between the pre-selection and final algorithm. The Spearman correlation represents a monotonic method that is more appropriate as a pre-selection method than the RF importance but the SVR importance represents the most consistent approach as the final algorithm is an SVR.

Variable Selection

The variables selected by the best performing model provide four diverse economic indicators that have a large explanatory power of systematic credit risk. The profitability expresses the ability of companies to generate capital, to service their debt. The profitability of non-financial companies is selected instead of all companies which is consistent with financial companies being filtered out of the portfolio and therefore, it is coherent that the non-financial variable delivers a higher accuracy. Further, the one-year lag of the variable

is selected, indicating that it takes some time between a reduction in profits to affect the ability to service debt. The commercial real estate price index presents a good proxy for companies expenses, with higher costs obviously limiting companies abilities to service their debt. Besides labour cost, the real estate cost represents a major component in a company's cost structure. The commercial real estate price falls when the demand for office and other commercial real estate space declines. This represents a cost sensitive indicator for companies' ability to service their debt as there is no lag selected. The BBB Spread is a measure for the risk premium of BBB rated corporate bonds and commonly used as a benchmark for financial markets' evaluation of corporate credit risk. It has a one year lag which shows the predictive power of financial markets. Finally, the US Export Goods variable represent the state of the exporting economy which indicates the competitiveness of the US economy and therefore the relative strength of US companies as well as the state of global trade. With the commercial real estate price index as the most impactful variable according to the variable importance measures and the profitability with the lowest, it is evident that the systematic credit risk is more sensitive to cost than to profit. Further, the US export goods volume, as a more general indicator for the economy, is a better gauge than the BBB spread, as the financial market view of systematic credit risk. The selected variables differ from the classical variables such as GDP growth and unemployment rate and more accurately describe the systematic credit risk. Compared with the risk factors, the Fed provides in its annual stress test, only the BBB Spread and commercial real estate price index are present. This justifies the extensive amount of variables considered and should encourage future researchers and model developers to consider a wider scope of variables.

Kernel and Hyperparameters

In this analysis, I will focus on the best performing model, the one resulting from the SVR importance pre-selection. The model is based on an RBF function in the variable selection with a gamma of 0.06 as shown in Table 4.2. The following hyperparameter tuning via cross-validation however, leads to a linear kernel function. This result as well as the

Model	Linear Benchmark	SVR Model	SVR Variables in OLS	Covariate Shift Model
OOS-MSE	0.683	0.3011	0.3267	0.27
OOS-MAE	0.6628	0.4464	0.4391	0.3546
OOS-adj.R^2	0	0.6154	0.5828	0.6510

Table 5.1: Out-of-Sample Performance Measures for Best Models

linear kernel functions of the best models with different pre-selection methods allows to conclude that the relationship between the macroeconomic variables and the Z-Factor is linear. With an epsilon of 0.5 the allowed margin of error is fairly large which means there is a wide channel around the fitted line where observations that lie in that channel are not penalized. However, the penalization for observations outside of the range is low with a C of 1. The impact these hyperparamters have on the model performance compared to a simple linear OLS model are shown in Table 5.1. Here the previous SVR results are displayed in column 3 and column 4 shows the results of the same four selected variables in a simple OLS regression. It is clear that these two model differ only slightly in respect to each of the measures with the SVR model performing marginally better except in the MAE. In terms of adjusted R^2 and MSE the SVR model performs 5.6% and 7.8% better than the OLS model. A small but significant outperformance is achieved with the use of SVR as the final model.

5.2 Linear Benchmark Model

In this section, I will discuss the results from the linear benchmark approach with a focus on the model in column 4 of Table 4.5. This model presents the best linear model in term of the performance measures and fulfills all of the Gauss-Markov assumptions.

5.2.1 Variable Selection

The best linear model consists of four variables. The US export volume is selected again as in the best machine learning model. As explained above it is an indicator for the

competitiveness of US economy and the state of global trade. Unlike in the machine learning model, the one-quarter lag of the variable is selected here which represent a minor difference and simply indicates the US export volume moves slightly ahead of the credit cycle. Further, the variables S&P Industrial and the Services PMI reflect two main sectors of the economy. The S&P Industrial represents the stock performance of the 73 largest listed industrial companies in the US and is a good gauge for the manufacturing sector of the economy. It has a one-quarter lag as well and thus, is slightly ahead of the credit cycle. The Services PMI is representative of the service sector in the economy. The nature of the variables, as a survey, is aiming to predict future economic outcomes. It is therefore not surprising that a long lag of one year is the best predictor for the Z-Factor since it is constructed as leading the business and credit cycle. Finally, Liquidity Swaps by the Fed with other central banks is among the selected variables, representing the impact of Fed actions on credit risk. However, there is no direct economic link as this balance sheet item does not directly affect the US market but rather international markets. An increase in liquidity swap lines can often be observed in times of financial distress and thus this variable shows a strong correlation with the Z-Factor but there is no direct causation. Again, the one-quarter lag of the variable is selected. It is noticeable that all of the variables have at least a lag of one quarter. The selection of the transformation of a variable is happening in the pre-selection step, which is done using the Pearson correlation as a criteria in this model and the cause for the different lag selection. With the Service PMI and Fed Liquidity Swaps with other central banks as the final variables selected, it is confirmed in the benchmark model that the consideration of a broad scope of variables is advantageous, to build an accurate model.

5.2.2 Performance Comparison with SVR Model

The benchmark model and the machine learning both return their best results for the variable set that is obtained through a transformation selection followed by VIF filtering. This confirms that this is the appropriate methodology to reduce the variable candidates. When

it comes to the model performance however, there are significant differences between the two models. While the in-sample performance of the linear model is good and better than the one of the SVR model, the large drop in the the out-of-sample performance for the linear model makes it significantly worse than the SVR model. The linear model displays a clear case of overfitting where the model loses flexibility to accurately predict new observations, due to a too tight fit in-sample. This leads to the SVR model substantially outperforming the linear model with 50% smaller out-of-sample prediction errors.

As previously shown in Table 5.1, by applying an OLS regression to the four variables of the best SVR model, the results only slightly differs from the SVR model results. This is evidence that the relationship between the macroeconomic variables is linear or at least close to linear. The large difference in performance between the linear benchmark and the machine learning model can therefore be attributed to the variable selection process. Here, the use of the SVR importance to select the transformation and the SVR algorithm used for the final variable selection make the difference compared to the linear variable selection approach using the Pearson correlation for the transformation selection and then the lasso regression. These selection methods are more appropriate to choose between the large amount of variables considered for this research since they avoid overfitting as observed in the benchmark model. The allowed margin of error is the mechanism in the SVR algorithm that allows for a more generalized mapping function between explanatory and independent variable. A more strict mapping function such as the Pearson correlation or the lasso regression result in variable selections that lead to overfitting models. As mentioned above the expansion of macroeconomic variables considered played an important part in the improvement of the model. Therefore, it is advisable to follow the variable selection approach with an allowed margin of error, even if the final model is a linear OLS model.

5.3 Covariate Shift Adaptation Model

In this section, I will discuss the results of the covariate shift adaptation on the three machine learning models presented in Table 4.6. The results of the implementation are discussed and interpreted with regard to the simple machine learning models.

5.4 Performance Comparison with SVR Model

The covariate shift adaptation requires a new hyperparameter optimization which leads to slightly different kernel functions but the majority remain linear, confirming the relationship between the macroeconomic variables and Z-Factor is linear. Further, the allowed margins of error are smaller and the penalization is higher for all three variable selections, suggesting a tighter fit of the mapping function to the observations. This is consistent with the use of sample weights in the covariate shift adaptation. The sample weights take into account the distribution of the out-of-sample observations which allows for the function to leave less room for observations to deviate which in turn is more accurate out-of-sample. This is evident by the significantly better out-of-sample performance across all three measures for the SVR pre-selection model which undoubtedly remains the best performing model. The covariate shift implementation here reduces the MSE by 10%, the MAE is reduced by 20% and the adjusted R^2 increases by 6%. Surprisingly, the Spearman correlation pre-selection model does not improve with the covariate shift implementation, despite the selected variables showing a large drift between train and test sample. The RF importance pre-selection model shows a slight improvement across the performance measures with the adjusted R^2 increasing from 0 to 0.06, while the MSE is reduced marginally by 1% and the MAE is reduced by 8%.

Conclusion

The approach presented in this study aims to improve the current stress testing methodology for corporate bonds. As Leo et al. (2019) mention, there is a lack of literature applying machine learning techniques to identify relationships between data, and use it for model selection and forecasting in stress testing. The presented approach offers a more accurate credit stress testing model, adapting a machine learning methodology.

The SVR machine learning model is clearly outperforming the linear benchmark model, due to the benchmark model strongly overfitting. The overfitting can be attributed to the variable selection process since the SVR model variables show a similar performance in an OLS regression as demonstrated in Table 5.1 and the relationship between the selected variables and the Z-Factor is shown to be linear. The characteristics of the SVR to allow for a margin of error in the regression, is better suited for the extensive list of macroeconomic variable candidates. This results in a 50% higher out-of-sample accuracy for the SVR model compared to the benchmark model. Both models profit from the extensive list of considered variables, as variables that are not common in the literature or proposed by the CCAR framework are selected in the best performing models. With the covariate shift adaptation which can account for the different distributions of the macroeconomic variables as evident between the training and test sample, the SVR model performance can further be improved.

With the obtained results the next step would be an application of the final model to a base and adverse economic scenario for the selected macroeconomic variables. However, since the NIPA profitability and Export Goods are not part of the stress testing scenarios

published by the Fed, a model would have to be constructed to obtain such scenarios. Once the Z-Factor is estimated with such a scenario, the PIT can be derived from the Z-Factor and with that the estimated default probability determined. Consequently, the impact of the shock on the portfolio's exposure at default and other measures can be obtained.

For future research, it would be interesting to extend the model application to the financial and the oil & gas portfolio. For these different sectors, most likely different variables would be selected for the best model, for example one of the Fed balance sheet variables could have a high predictive power for the financial portfolio. Energy related variables such as the WTI price could be an important variable in the oil & gas portfolio. Another possible research topic would be to estimate models of the corporate portfolio on a more granular level as done by Chan-Lau (2017) for the probability of default prediction. This could mean dividing the portfolio into industry sectors according to Table 2.2 and using sector specific variables for the Z-Factor estimation. This has the potential for a more accurate estimation of sector specific credit risks but a challenge that arises with such an approach is data scarcity. A large number of obligors in each sector is required to obtain stable transition matrices for the Z-Factor extraction and this is not the case for every sector.

Further, a longer observation period can confirm the selected model and therefore, testing the model on a dataset with a larger history can provide further insights. The impact, the selection of the test and train time frame can have, becomes evident if solely the last two observations, Q1 and Q2 of 2020 are removed. The modified test set leads to a reduction in adjusted R^2 by almost 50% compared to the previous SVR model. This shows the sensitivity of the results to the selection of the test and training dataset and the need for a validation of the results over a longer time frame.

The extracted Z-Factor displays first order autocorrelation as evident by the ACF and PACF in Figure A22 in the Appendix. This is well known in the credit rating literature as rating drift. As Bandt et al. (2013) show, an incorporation of this characteristic could further improve the accuracy of the model. A possibility to integrate the autocorrelation

in the model presented here, could be a hybrid ARIMA SVR model as proposed by Pai and Lin (2005).

Bibliography

Bandt, O. D., Dumontaux, N., Martin, V., and Médée, D. (2013). Stress-testing banks' corporate credit portfolio. Débats économiques et financiers 1, Banque de France.

Bangia, A., Diebold, F. X., Kronimus, A., Schagen, C., and Schuermann, T. (2002). Ratings migration and the business cycle, with application to credit portfolio stress testing. *Journal of Banking & Finance*, 26(2):445 – 474.

Basak, D., Pal, S., and Patranabis, D. (2007). Support vector regression. *Neural Information Processing–Letters and Reviews*, 11(10):203–224.

Belkin, B., Suchower, S., and Forest, L. R. (1998a). The effect of systematic credit risk on loan portfolio value-at-risk and loan pricing. *CreditMetrics Monitor*, pages 17–28.

Belkin, B., Suchower, S., and Forest Jr, L. (1998b). A one-parameter representation of credit risk and transition matrices. *CreditMetrics monitor*, 1(3):46–56.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Chan-Lau, M. J. A. (2017). *Lasso regressions and forecasting models in applied stress testing*. International Monetary Fund.

Cherkassky, V. and Ma, Y. (2004). Practical selection of svm parameters and noise estimation for svm regression. *Neural Networks*, 17(1):113–126.

Figlewski, S., Frydman, H., and Liang, W. (2012). Modeling the effect of macroeconomic factors on corporate default and credit rating transitions. *International Review of Economics & Finance*, 21(1):87 – 105.

Guajardo, J., Weber, R., and Miranda, J. (2006). A forecasting methodology using support vector regression and dynamic feature selection. *Journal of Information & Knowledge Management*, 05(04):329–335.

Hajek, P. and Michalak, K. (2013). Feature selection in corporate credit rating prediction. *Knowledge-Based Systems*, 51:72 – 84.

Harrell, F. E. (2017). Regression modeling strategies. *BIOS*, 330:2018.

Hayden, R. W. (2005). A review of: "applied linear regression models". *Journal of Biopharmaceutical Statistics*, 15(3):531–533.

Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H., and Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems*, 37(4):543 – 558. Data mining for financial decision making.

Jacobs Jr, M. (2018). The validation of machine-learning models for the stress testing of credit risk. *Journal of Risk Management in Financial Institutions*, 11(3):218–243.

Kadam, A. and Lenk, P. (2008). Bayesian inference for issuer heterogeneity in credit ratings migration. *Journal of Banking & Finance*, 32(10):2267 – 2274.

Lando, D. and Skødeberg, T. M. (2002). Analyzing rating transitions and rating drift with continuous observations. *Journal of Banking & Finance*, 26(2-3):423–444.

Lee, Y.-C. (2007). Application of support vector machines to corporate credit rating prediction. *Expert Systems with Applications*, 33(1):67 – 74.

Leo, M., Sharma, S., and Maddulety, K. (2019). Machine learning in banking risk management: A literature review. *Risks*, 7(1):29.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, 29(2):449–470.

Nickell, P., Perraudin, W., and Varotto, S. (2000). Stability of rating transitions. *Journal of Banking & Finance*, 24(1):203 – 227.

Pai, P.-F. and Lin, C.-S. (2005). A hybrid arima and support vector machines model in stock price forecasting. *Omega*, 33(6):497 – 505.

Schuermann, T. and Hanson, S. G. (2004). Estimating probabilities of default. Staff Report 190, New York, NY.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317.

Sugiyama, M. and Kawanabe, M. (2012). Machine learning in non-stationary environments : introduction to covariate shift adaptation.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Vapnik, V., Golowich, S. E., and Smola, A. J. (1997). Support vector method for function approximation, regression estimation and signal processing. In *Advances in neural information processing systems*, pages 281–287.

Vasicek, O. A. (1987). *Probability of loss on loan portfolio*. KMV.

Yao, X., Crook, J., and Andreeva, G. (2015). Support vector regression for loss given default modeling. *European Journal of Operational Research*, 240(2):528–538.

Zhang, H., Nettleton, D., and Zhu, Z. (2019). Regression-enhanced random forests.

Appendix

Table A1: Spearman Correlation Pre-Selection

Pre VIF Filter	Post VIF Filter
US export goods vol QGrowth	US export goods vol QGrowth
DSR Corps non financial QGrowth Lag12m	DSR Corps non financial QGrowth Lag12m
Commercial RE PI QGrowth	Commercial RE PI QGrowth
BBB Spread QD Lag12m	BBB Spread QD Lag12m
NIPA Profitability non financial YGrowth Lag3m	S&P Industrial YGrowth Lag3m
Nominal GDP Growth YD Lag12m	NIPA Profitability non financial YGrowth Lag3m
S&P Industrial YGrowth Lag3m	US Treasury general account YGrowth
NIPA Profitability QGrowth Lag12m	ISM PMI Manufacturing YGrowth Lag6m
Total Securities QGrowth	Initial Claims QGrowth Lag9m
ISM PMI Manufacturing YGrowth Lag6m	NIPA Profitability QGrowth Lag12m
Loans YGrowth Lag3m Total	Securities QGrowth
US Treasury general account YGrowth	VIX YGrowth Lag3m
ISM PMI Services YGrowth Lag12m	Loans YGrowth Lag3m
VIX YGrowth Lag3m	Nominal GDP Growth YD Lag12m
M2 Velocity QGrowth Lag9m	Real GDP Growth Lag3m
Total Assets YGrowth	M2 Velocity QGrowth Lag9m
Lending TS LM Corps Lag9m	Liq. Swaps w/ CBs Lag3m
Liq. Swaps w/ CBs Lag3m	Lending TS ConsCC Lag9m
Initial Claims QGrowth Lag9m	ISM PMI Services YGrowth Lag12m
DJ Total SM Index YGrowth Lag3m	
Lending TS ConsCC Lag9m	
Real GDP Growth Lag3m	

Table A2: Variables Selected with RF

Pre VIF Filter	Post VIF Filter
Loans QGrowth Lag12m	Loans QGrowth Lag12m
BBB Spread QD Lag12m	S&P Industrial YGrowth Lag3m
US export goods vol YGrowth Lag3m	BBB Spread QD Lag12m
Commercial RE PI QGrowth	DSR Corps non financial QGrowth Lag12m
DSR Corps non financial QGrowth Lag12m	US export goods vol YGrowth Lag3m
ISM PMI Services YGrowth Lag12m	Commercial RE PI QGrowth
NIPA Profitability QGrowth Lag12m	NIPA Profitability QGrowth Lag12m
S&P Industrial YGrowth Lag3m	Real GDP Growth YD Lag6m
NIPA Profitability non financial YGrowth Lag3m	ISM PMI Services YGrowth Lag12m
Initial Claims QGrowth Lag6m	Initial Claims QGrowth Lag6m
M2 Velocity QGrowth Lag9m	NIPA Profitability non financial YGrowth Lag3m
Real GDP Growth YD Lag6m	M2 Velocity QGrowth Lag9m
DJ Total SM Index YGrowth Lag3m	Total Securities QGrowth
Total Securities QGrowth	Recession Dummy
ISM PMI Manufacturing YGrowth Lag6m	US Treasury general account YGrowth Lag3m
Recession Dummy	Total Assets YGrowth
Total Assets YGrowth	Lending TS ConsCC Lag9m
Lending TS ConsCC Lag9m	VIX QGrowth Lag6m
Nominal GDP Growth YD Lag6m	Lending TS LM Corps Lag9m
US Treasury general account YGrowth Lag3m	Liq. Swaps w/ CBs Lag3m
Lending TS LM Corps Lag9m	ISM PMI Manufacturing YGrowth Lag6m
VIX QGrowth Lag6m	
Liq. Swaps w/ CBs Lag3m	

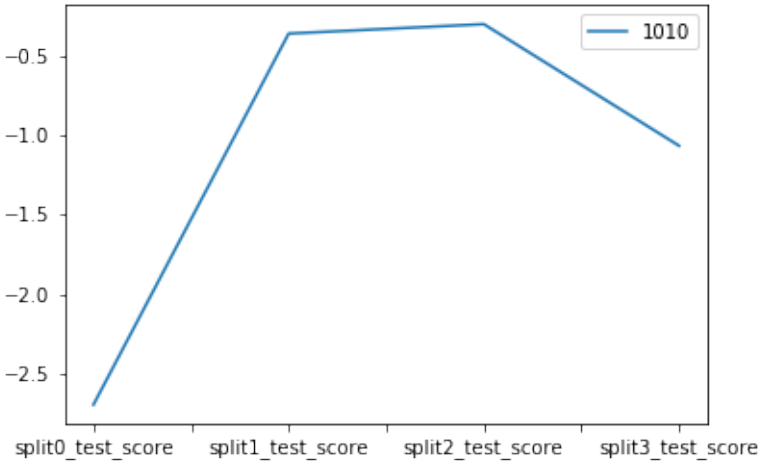
Table A3: Support Vector Regression Pre-Selection

Pre VIF Filter	Post VIF Filter
BBB Spread QD Lag12m	BBB Spread QD Lag12m
NIPA Profitability non financial QGrowth Lag12m	NIPA Profitability non financial QGrowth Lag12m
US export goods vol QGrowth	ISM PMI Manufacturing YGrowth Lag3m
Commercial RE PI QGrowth	DSR Corps non financial QGrowth Lag12m
ISM PMI Manufacturing YGrowth Lag3m	M2 Velocity QGrowth Lag12m
DSR Corps non financial QGrowth Lag12m	Commercial RE PI QGrowth
ISM PMI Services YGrowth Lag12m	NIPA Profitability YGrowth Lag3m
NIPA Profitability YGrowth Lag3m	S&P Industrial YGrowth Lag3m
M2 Velocity QGrowth Lag12m	US Treasury general account QGrowth Lag3m
US Treasury general account QGrowth Lag3m	US export goods vol QGrowth
Initial Claims QGrowth	Initial Claims QGrowth
Recession Dummy	Total Securities QGrowth
DJ Total SM Index YGrowth Lag3m	Lending TS ConsCC Lag9m
S&P Industrial YGrowth Lag3m	ISM PMI Services YGrowth Lag12m
Total Securities QGrowth	Lending TS LM Corps Lag9m
Lending TS ConsCC Lag9m	Nominal GDP Growth Lag3m
Loans Lag3m	Recession Dummy
Total Assets QGrowth Lag3m	VIX QGrowth Lag6m
Liq. Swaps w/ CBs Lag3m	
Lending TS LM Corps Lag9m	
Real GDP Growth Lag3m	
Nominal GDP Growth Lag3m	
VIX QGrowth Lag6m	

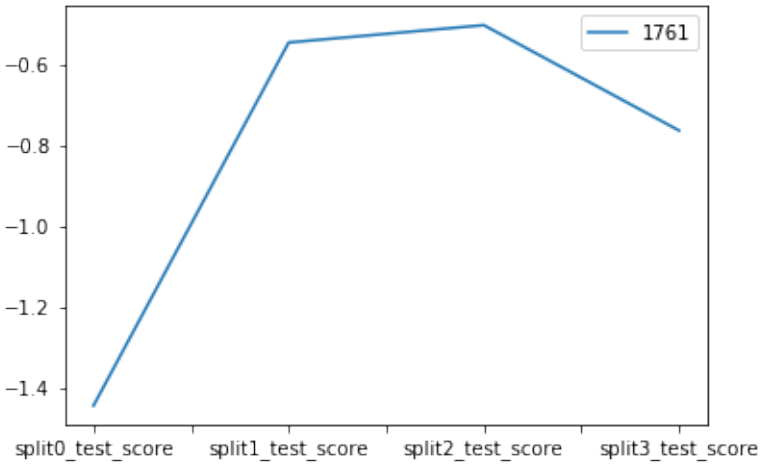
Table A4: Importance Measures for Variable Selection in RF Pre-Selection Model before VIF Filter

Permutation Importance		Drop-Column Importance		Shapley Value	
S&P Industrial YGrowth Lag3m	0.09	Loans QGrowth Lag12m	0.03	NIPA Profitability QGrowth Lag12m	0.10
DSR Corps non financial QGrowth Lag12m	0.09	BBB Spread QD Lag12m	0.01	DSR Corps non financial QGrowth Lag12m	0.09
BBB Spread QD Lag12m	0.09	US export goods vol YGrowth Lag3m	0.01	S&P Industrial YGrowth Lag3m	0.09
NIPA Profitability QGrowth Lag12m	0.08	Commercial RE PI QGrowth	0.01	US export goods vol YGrowth Lag3m	0.08
Loans QGrowth Lag12m	0.08	DSR Corps non financial QGrowth Lag12m	0.01	BBB Spread QD Lag12m	0.08
US export goods vol YGrowth Lag3m	0.08	ISM PMI Services YGrowth Lag12m	0.01	Loans QGrowth Lag12m	0.07
Liq. Swaps w/ CBs Lag3m	0.08	NIPA Profitability QGrowth Lag12m	0.01	ISM PMI Manufacturing YGrowth Lag6m	0.07
ISM PMI Manufacturing YGrowth Lag6m	0.06	S&P Industrial YGrowth Lag3m	0.01	Nominal GDP Growth YD Lag6m	0.06
Commercial RE PI QGrowth	0.06	NIPA Profitability non financial YGrowth Lag3m	0.00	Commercial RE PI QGrowth	0.06
Nominal GDP Growth YD Lag6m	0.05	Initial Claims QGrowth Lag6m	0.00	NIPA Profitability non financial YGrowth Lag3m	0.06

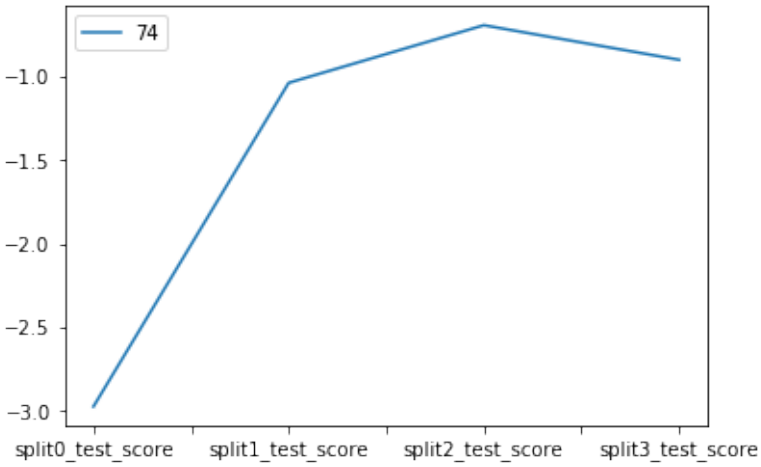
Figure A1: Cross-Validation Results for best Spearman Correlation Model



(a) MSE



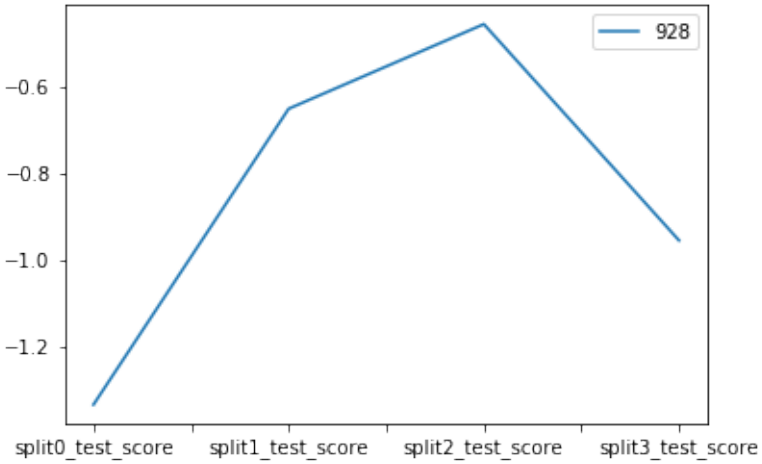
(b) MAE



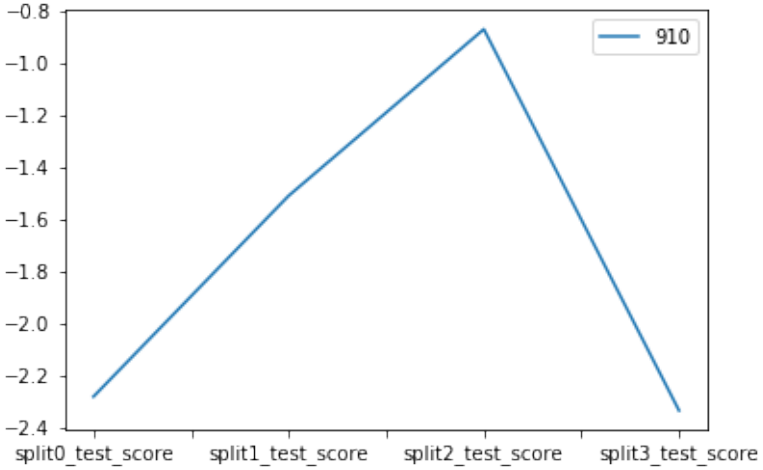
(c) R^2

Figure A2: Cross-Validation Results for best RF Model

(a) MSE

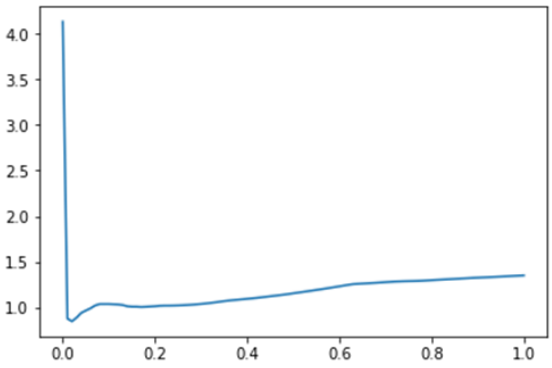


(b) MAE

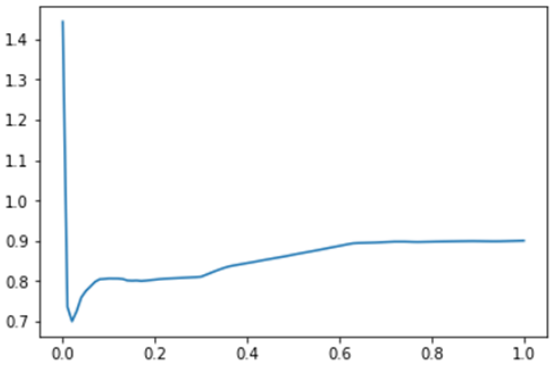


(c) R^2

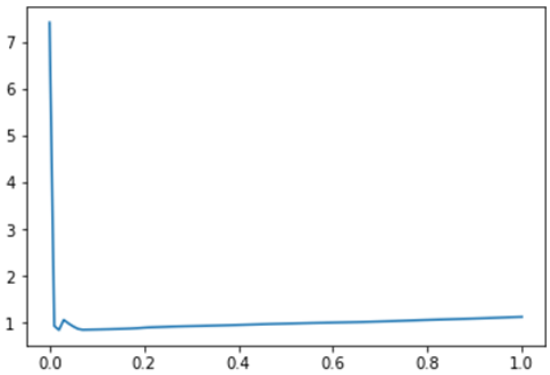
Figure A3: Cross-Validation Results for lambda Selection in Lasso



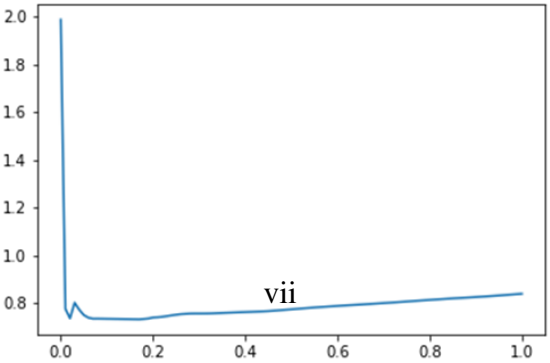
(a) VIF Filtered Set - MSE



(b) VIF Filtered Set - MAE



(c) Pearson Correlation Set - MSE



vii

(d) Pearson Correlation Set - MSE

Table A5: Importance Measures for Variable Selection in RF Pre-Selection Model after VIF Filter

Permutation Importance		Drop-Column Importance		Shapley Value	
S&P Industrial YGrowth Lag3m	0.11	Loans QGrowth Lag12m	0.03	S&P Industrial YGrowth Lag3m	0.11
BBB Spread QD Lag12m	0.10	S&P Industrial YGrowth Lag3m	0.01	US export goods vol YGrowth Lag3m	0.09
Liq. Swaps w/ CBs Lag3m	0.09	BBB Spread QD Lag12m	0.01	BBB Spread QD Lag12m	0.09
Loans QGrowth Lag12m	0.09	DSR Corps non financial QGrowth Lag12m	0.01	NIPA Profitability QGrowth Lag12m	0.09
US export goods vol YGrowth Lag3m	0.08	US export goods vol YGrowth Lag3m	0.01	Real GDP Growth YD Lag6m	0.09
DSR Corps non financial QGrowth Lag12m	0.08	Commercial RE PI QGrowth	0.01	DSR Corps non financial QGrowth Lag12m	0.08
NIPA Profitability QGrowth Lag12m	0.08	NIPA Profitability QGrowth Lag12m	0.01	Loans QGrowth Lag12m	0.08
Real GDP Growth YD Lag6m	0.08	Real GDP Growth YD Lag6m	0.01	Commercial RE PI QGrowth	0.07
Commercial RE PI QGrowth	0.07	ISM PMI Services YGrowth Lag12m	0.01	ISM PMI Services YGrowth Lag12m	0.06
US Treasury general account YGrowth Lag3m	0.06	Initial Claims QGrowth Lag6m	0.00	NIPA Profitability non financial YGrowth Lag3m	0.06

Table A6: Spearman Correlation Pre-Selection Model Performance before and after VIF Filter

Transformation Selection	Pre VIF Filter	Post VIF Filter
Variables	US export goods vol QGrowth, DSR Corps non financial QGrowth Lag12m, BBB Spread QD Lag12m	US export goods vol QGrowth, DSR Corps non financial QGrowth Lag12m, BBB Spread QD Lag12m
Hyperparameters	Kernel=sigmoid, C=3, ε=0.1, γ=0.001 Kernel=sigmoid, C=3, ε=0.1, γ=0.001 Kernel=rbf, C=2, ε=0.1, γ=0.001	Kernel=sigmoid, C=3, ε=0.1, γ=0.001 Kernel=sigmoid, C=3, ε=0.1, γ=0.001 Kernel=rbf, C=2, ε=0.1, γ=0.001
CV-MSE	1.0340	1.0340
CV-MAE	0.7258	0.7258
CV-adj. R^2	0.174	0.174
OOS Performance - Candidate Model 1		
MSE	0.4	0.4
MAE	0.5	0.5
adj. R^2	0.49	0.49
OOS Performance - Candidate Model 2		
MSE	0.43	0.43
MAE	0.50	0.5
adj. R^2	0.46	0.46

Table A7: RF Importance Pre-Selection Model Performance before and after VIF Filter

Transformation Selection	Pre VIF Filter	Post VIF Filter
Variables	NIPA Profitability QGrowth Lag12m, DSR Corps non financial QGrowth Lag12m, BBB Spread QD Lag12m, Loans QGrowth Lag12m	US export goods vol YGrowth Lag3m, BBB Spread QD Lag12m, S&P Industrial YGrowth Lag3m
Hyperparameters	Kernel=sigmoid, C=1, ε=0.6, γ=0.0001 Kernel=sigmoid, C=1, ε=0.8, γ=0.0001 Kernel=sigmoid, C=1, ε=0.6, γ=0.0001	Kernel=linear, C=2, ε=0.4 Kernel=linear, C=2, ε=0.4 Kernel=linear, C=2, ε=0.4
CV-MSE	3.9498	0.4628
CV-MAE	0.8581	0.5750
CV-adj.R^2	-2.1501	0.6308
OOS Performance - Candidate Model 1		
MSE	0.71	0.8
MAE	0.71	0.74
adj.R^2	0	0
OOS Performance - Candidate Model 2		
MSE	0.74	
MAE	0.93	
adj. R^2	0	

Table A8: SVR Importance Pre-Selection Model Performance before and after VIF Filter

Transformation Selection	Pre Vif Filter	Post VIF Filter
Variables	US export goods vol QGrowth, NIPA Profitability non financial QGrowth Lag12m	NIPA Profitability non financial QGrowth Lag12m, Commercial RE PI QGrowth, BBB Spread QD Lag12m, US export goods vol QGrowth
Hyperparameters	Kernel=rbf, C=5, ε=0.5, γ=0.001 Kernel=rbf, C=5, ε=0.5, γ=0.001 Kernel=rbf, C=5, ε=0.5, γ=0.001	Kernel=linear, C=1, ε=0.5 Kernel=linear, C=1, ε=0.5 Kernel=linear, C=1, ε=0.5
CV-MSE	0.8384	0.4579
CV-MAE	0.6494	0.5382
CV-adj.R^2	0.3223	0.6348
OOS Performance - Candidate Model 1		
MSE	0.34	0.3
MAE	0.43	0.45
adj. R^2	0.57	0.62

Table A9: Linear Model Performance for VIF filtered Pre-selection

Transformation Selection	Variables VIF Filtered Variables	VIF Filtered & significant	Reduction to 4 Variables	Reduction to 3 Variables
Variables	No variable removes (22)	S&P Industrial QGrowth Lag3m, S&P Industrial QGrowth Lag6m, S&P Industrial QGrowth Lag9m, ... (11)	S&P Industrial QGrowth Lag9m, VIX Ygrowth Lag3m, VIX Ygrowth Lag12m, Real GDP Growth YD Lag6m	S&P Industrial QGrowth Lag9m, VIX Ygrowth Lag12m, Real GDP Growth YD Lag6m
Lambda	0.02	0.02	0.5	0.55
IS-MSE	0.0954	0.142	0.4181	0.4851
IS-MAE	0.2265	0.3055	0.4922	0.5647
IS-adj.R^2	0.8772	0.8102	0.4621	0.3758
OOS-MSE	1.033	1.059	1.121	1.111
OOS-MAE	0.9438	0.9391	0.9008	0.8883
OOS-R^2 adj.	0	0	0	0

Table A10: Best Spearman Correlation Model Importance Measures

Variable	Permutation Importance	Drop-Column Importance	Shapley Value
BBB Spread QD Lag12m	0.04	0.04	0.10
DSR Corps non financial QGrowth Lag12m	0.03	0.01	0.12
US export goods vol QGrowth	0.02	0.02	0.11

Table A11: Best RF Model Importance Measures

Variable	Permutation Importance	Drop-Column Importance	Shapley Value
BBB Spread QD Lag12m	0.001	0.001	0.11
NIPA Profitability QGrowth Lag12m	0.0006	0.0007	0.08
Loans QGrowth Lag12m	0.0007	0.0006	0.025
DSR Corps non financial QGrowth Lag12m	0.0003	0.0003	0.09

Figure A4: In-Sample Performance of best Spearman Correlation Model

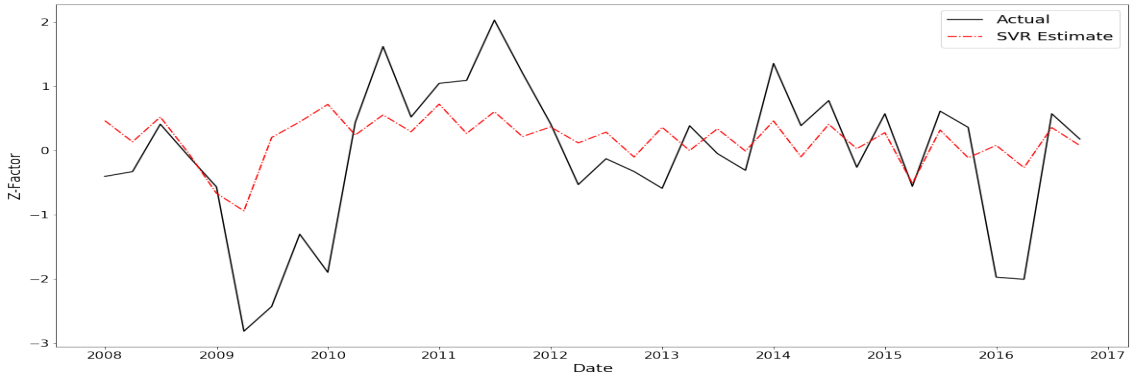


Figure A5: Out-of-Sample Performance of best Spearman Correlation Model

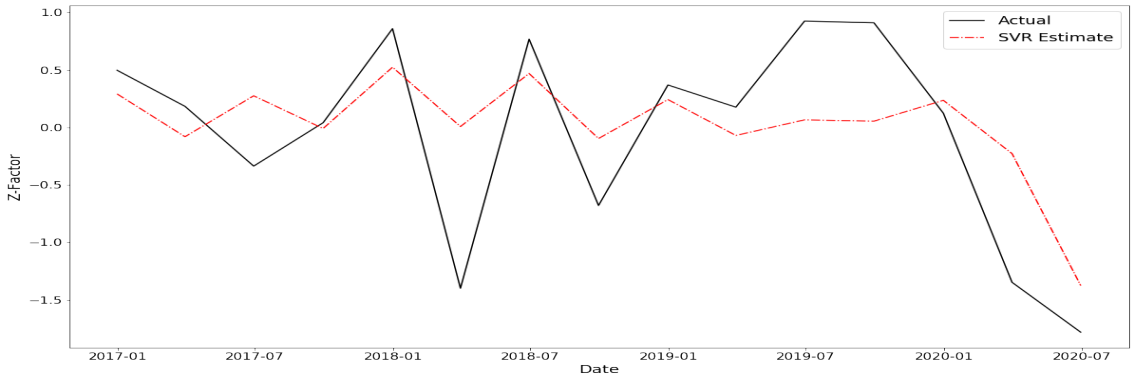


Figure A6: In-Sample Performance of best Spearman Correlation Model with MEVs

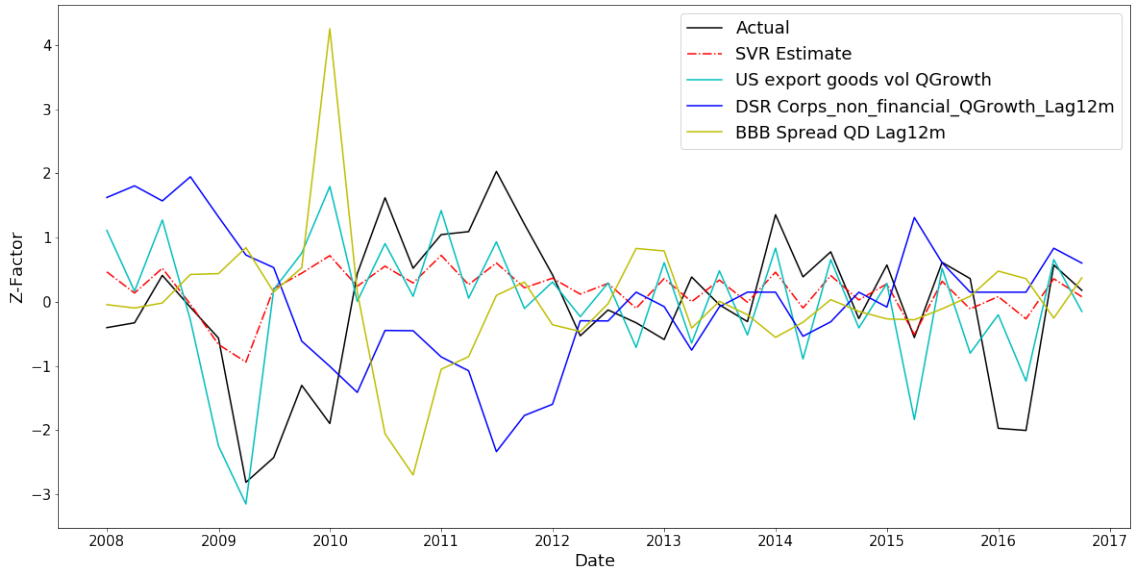


Figure A7: Out-of-Sample Performance of best Spearman Correlation Model with MEVs

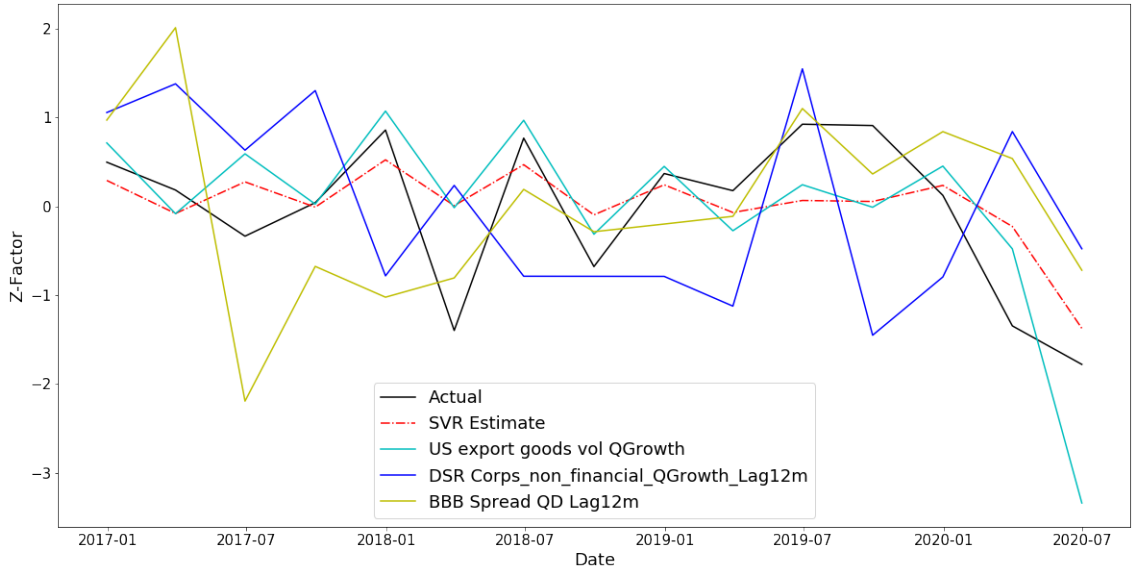


Figure A8: In-Sample Performance of best RF Model

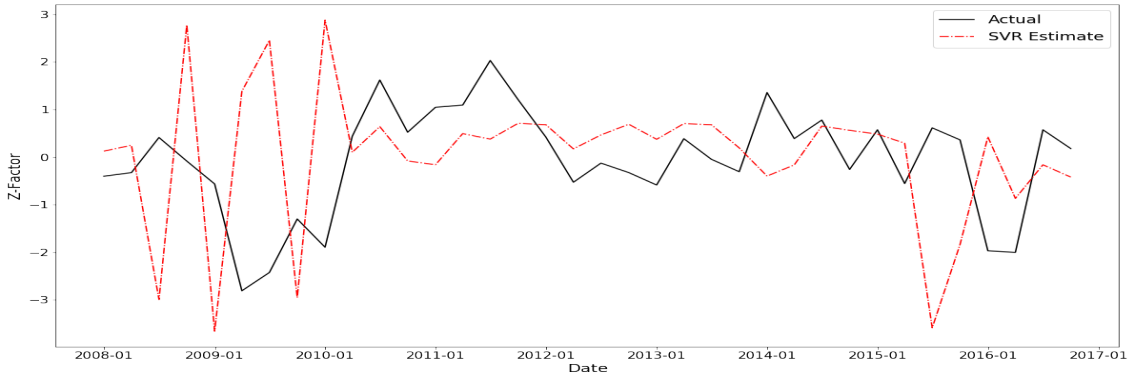


Figure A9: Out-of-Sample Performance of best RF Model

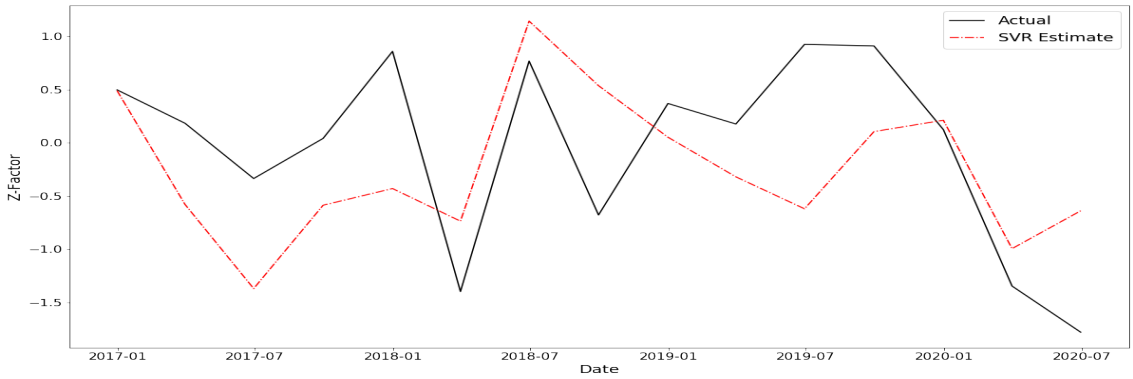


Figure A10: In-Sample Performance of best RF Model with MEVs

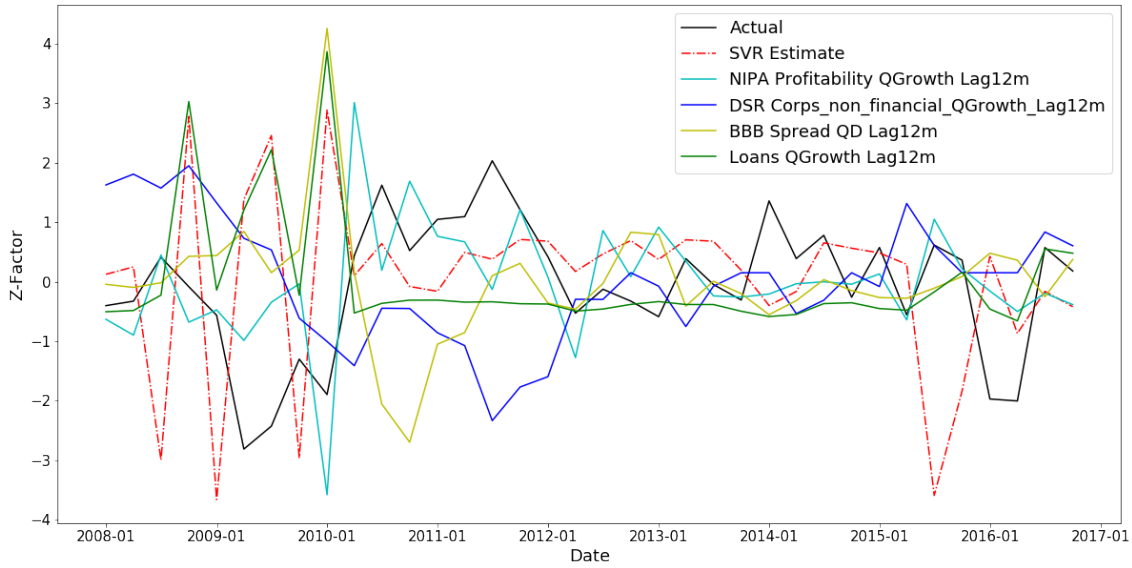


Figure A11: Out-of-Sample Performance of best RF Model with MEVs

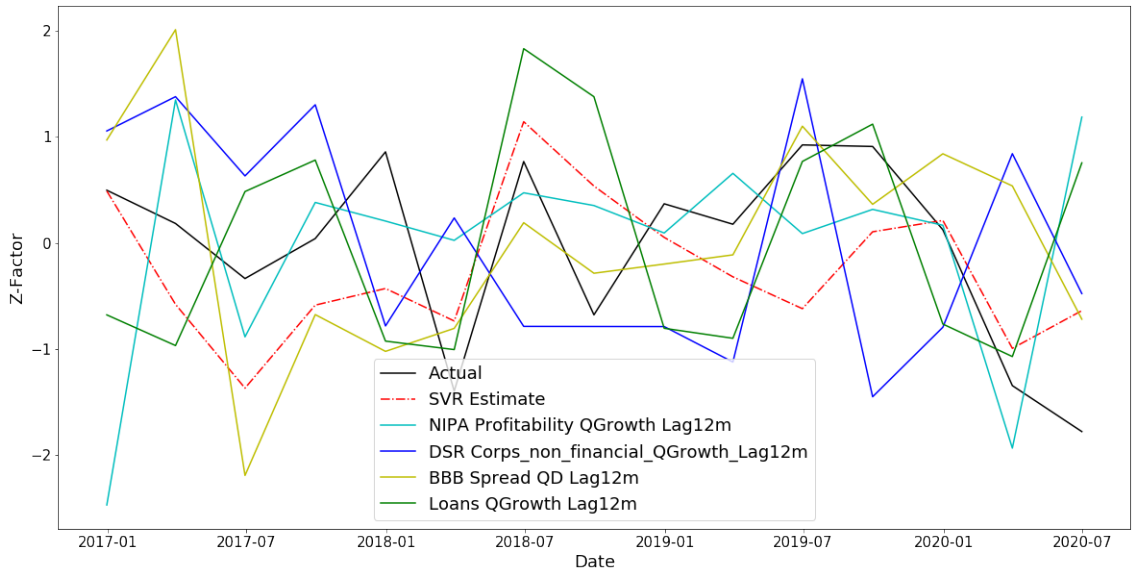


Figure A12: In-Sample Performance of best Benchmark Model

Figure A13: Out-of-Sample Performance of best Benchmark Model

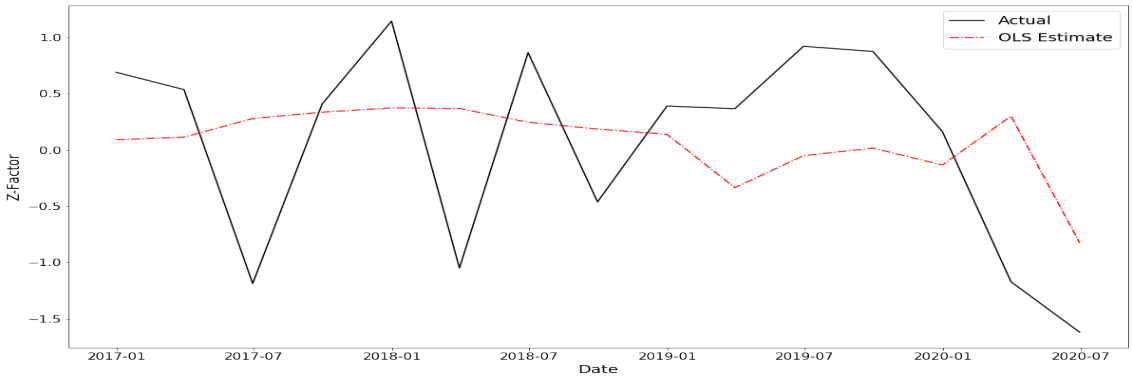


Figure A14: In-Sample Performance of best Benchmark Model with MEVs

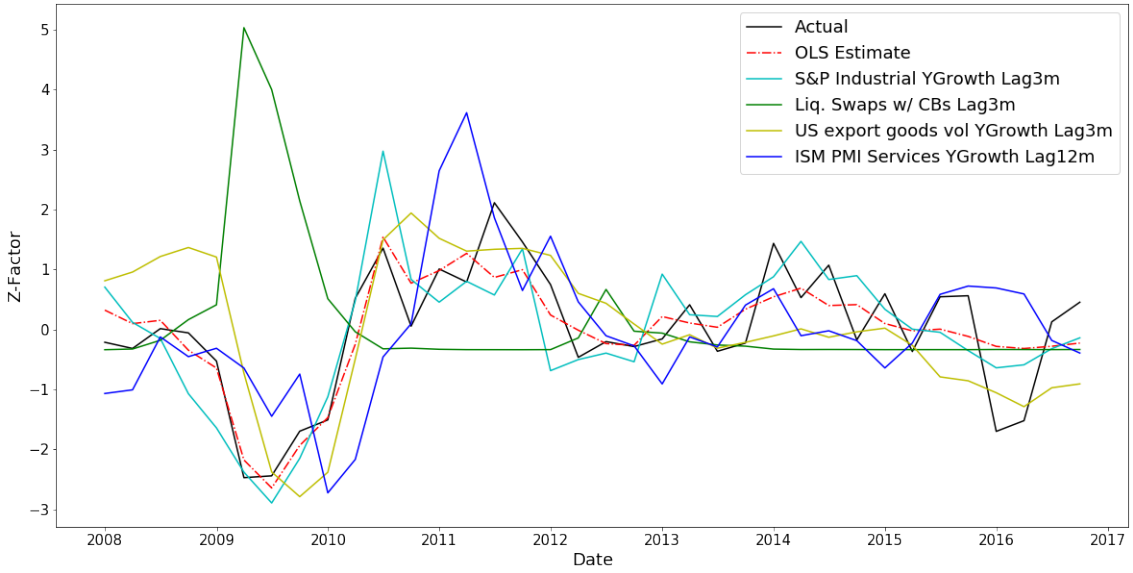


Figure A15: Out-of-Sample Performance of best Benchmark Model with MEVs

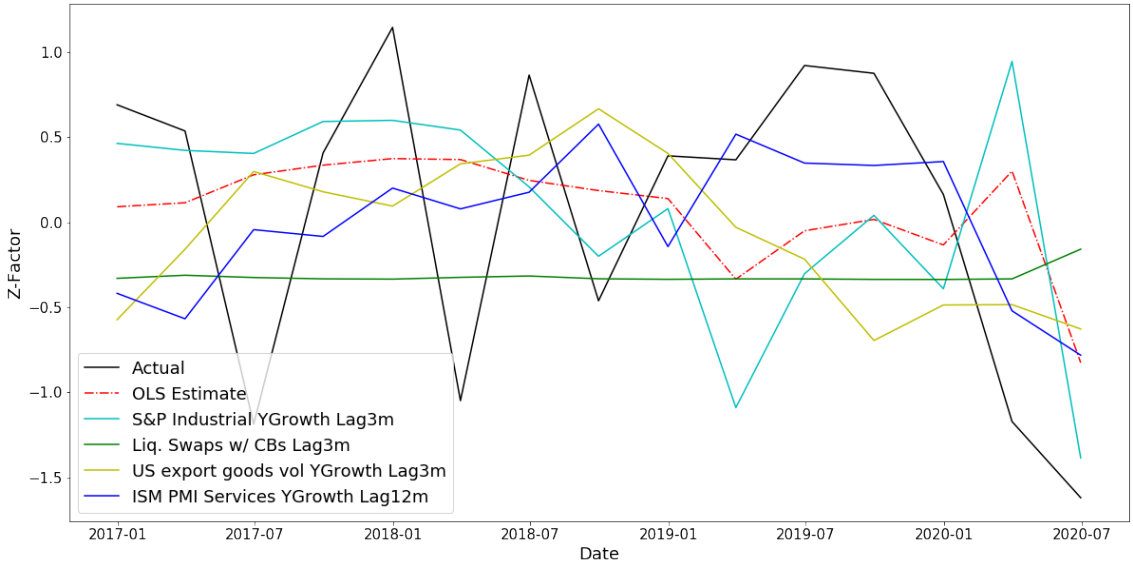


Figure A16: In-Sample Performance of best Covariate Shift Model

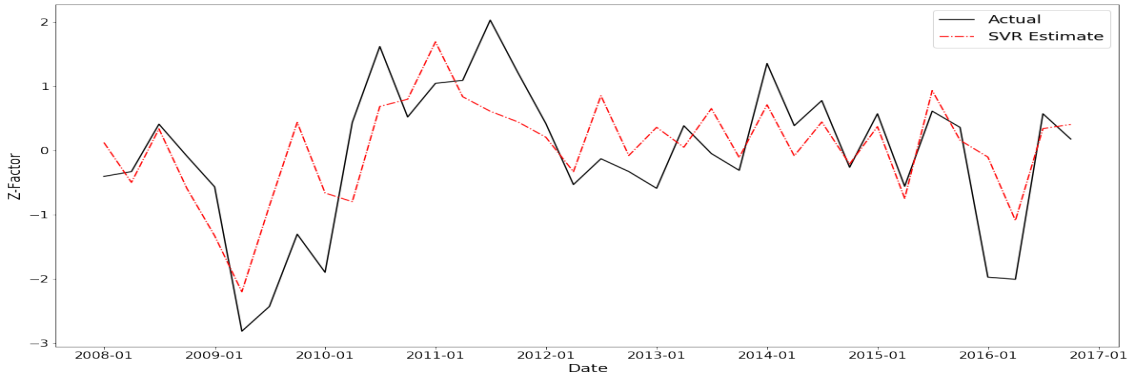


Figure A17: Out-of-Sample Performance of best Covariate Shift Model

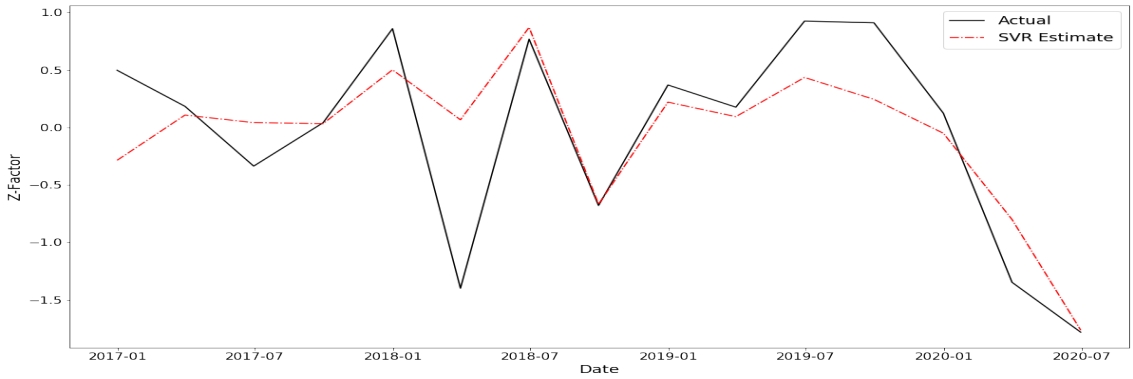


Figure A18: In-Sample Performance of best Covariate Shift Model with MEVs

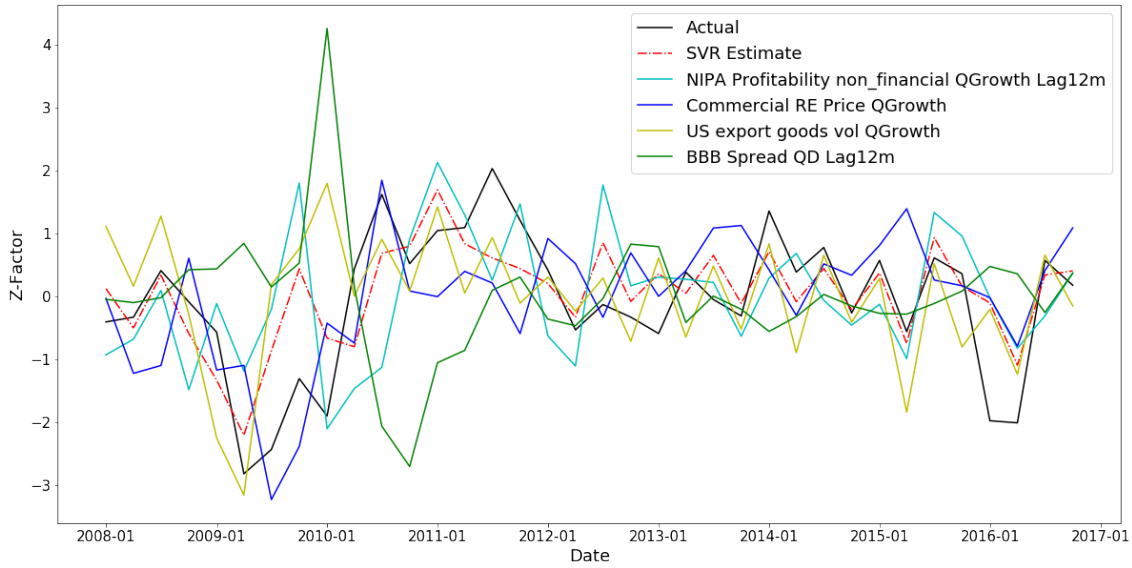


Figure A19: Out-of-Sample Performance of best Covariate Shift Model with MEVs

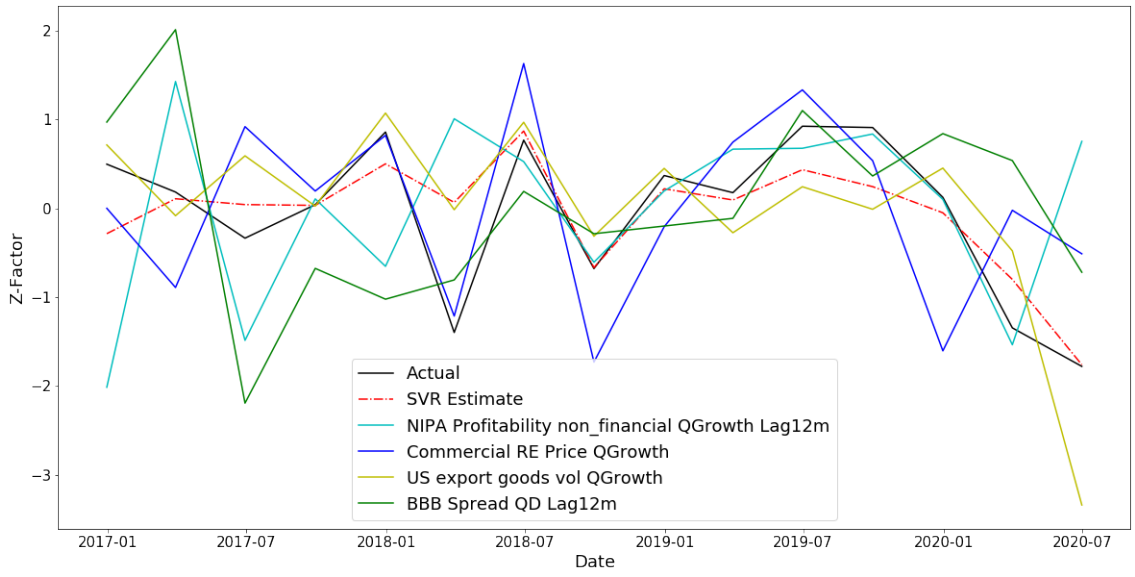
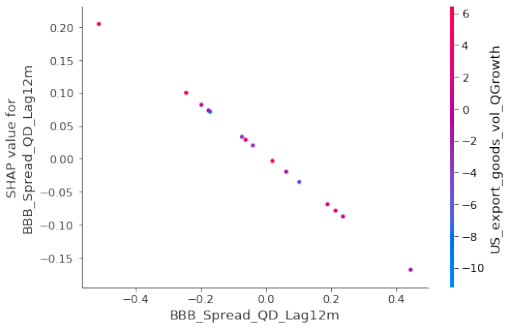
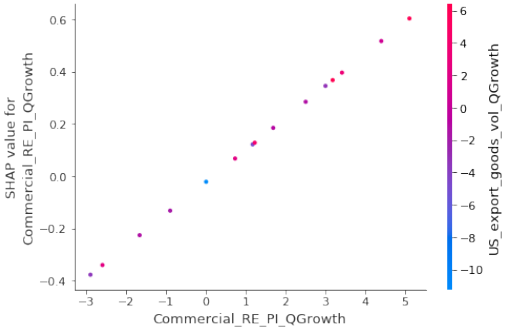


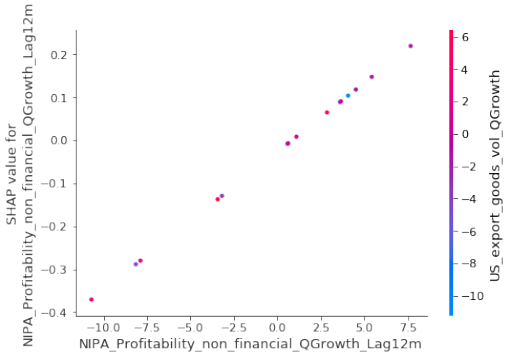
Figure A20: Dependency Plots for best SVR Model



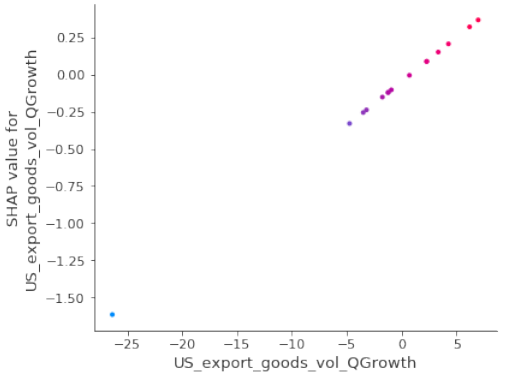
(a) BBB Spread



(b) Commercial RE PI



(c) NIPA non-financial Profitability



(d) US Export Goods Volume

Figure A21: Sample Weights for SVR Covariate Shift Model

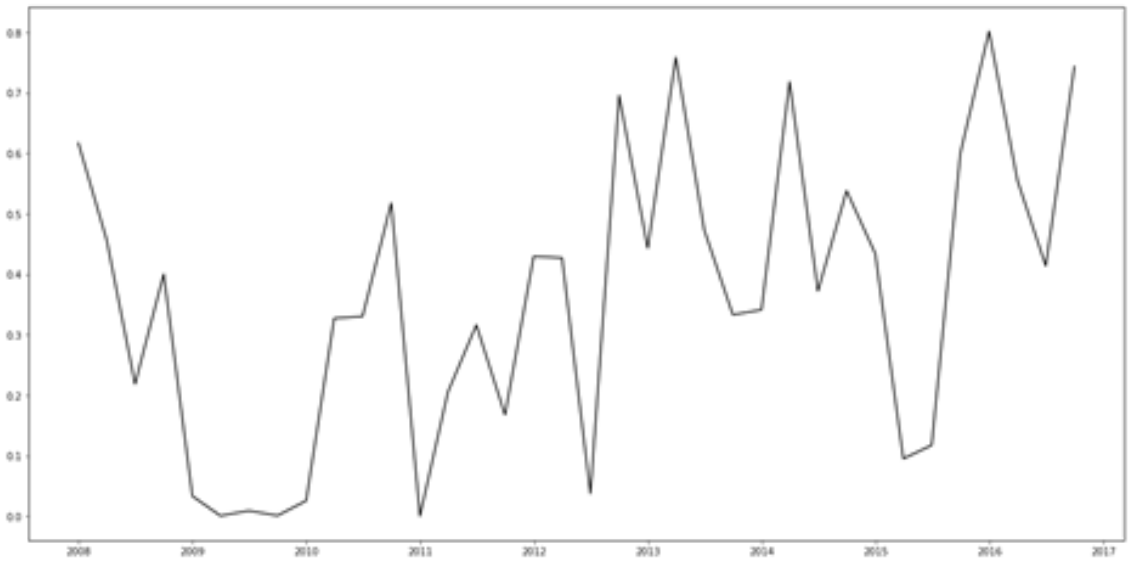


Table A12: Gauss-Markov Assumptions for VIF Filtered Variable Set

Transformation Selection	VIF Filtered	VIF Filtered & significant	Reduction to 4 variables	Reduction to 3 Variables
Normality of residuals (Jarque Bera Test, SW)	JB p-Value =0.67 ✓, SW p-value = 0	JB p-Value =0.95 ✓, SW p-value = 0	JB p-Value =0.94 ✓, SW p-value = 0.002	JB p-Value =0.89 ✓, SW p-value = 0.03
Multicollinearity (VIFs)	Multiple VIFs above 5	All VIFs below 5 ✓	All VIFs below 5 ✓	All VIFs below 5 ✓
Homoskedasticity (Breusch-Pagan & Goldfeld-Quandt Test)	BP P-Value = 0.99 ✓, GQ P-Value = nan	BP P-Value = 0.80 ✓, GQ P-Value = 0.19 ✓	BP P-Value = 0.34 ✓, GQ P-Value = 0.25 ✓	BP P-Value = 0.26 ✓, GQ P-Value = 0.46 ✓
Autocorrelation of Residuals (Durbin Watson Test)	Test statistic = 2.69	Test statistic = 2.32✓	Test statistic = 1.47 ✓	Test statistic = 1.47 ✓

Table A13: Gauss-Markov Assumptions for VIF Filtered & Pearson Correlation Selection Variable Set

Transformation Selection	VIF Filtered	VIF Filtered & significant	Reduction to 4 variables	Reduction to 3 Variables
Normality of residuals (Jarque Bera Test)	JB p-Value =0.38 ✓, SW p-value = 0.00	JB p-Value =0.65 ✓, SW p-value = 0.1 ✓	JB p-Value =0.38 ✓, SW p-value = 0.00	JB p-Value =0.79 ✓, SW p-value = 0.00
Multicollinearity (VIFs)	All VIFs below 5 ✓	All VIFs below 5 ✓	All VIFs below 5 ✓	All VIFs below 5 ✓
Homoskedasticity (Breusch-Pagan & Goldfeld-Quandt Test)	BP P-Value = 0.24 ✓, GQ P-Value = 0.08 ✓	BP P-Value = 0.26 ✓, GQ P-Value = 0.25 ✓	BP P-Value = 0.15 ✓, GQ P-Value = 0.08 ✓	BP P-Value = 0.35 ✓, GQ P-Value = 0.2 ✓
Autocorrelation of Residuals (Durbin Watson Test)	Test statistic = 2.03 ✓	Test statistic = 2.11 ✓	Test statistic = 2.16 ✓	Test statistic = 2.05 ✓

Figure A22: Z-Factor Autocorrelation