Erasmus School of Economics

Quantitative Finance

Crisis Adjustments to Credit Risk Models

Student

Stèphan Vegter (430628)

Supervisors

Dr. A.A. Naghi

B. van Oers (Zanders)

February 24, 2022

Abstract

The paper aims to investigate whether the performance of credit risk models has decreased as a consequence of the COVID-19 pandemic. It has the focus is on the credit risk models in The Netherlands and uses historical financial data of a wide range of companies. A Logistic Regression model is used as a benchmark model after which a Panel Smooth Transition Regression model is created and tested to see whether it can outperform the benchmark model. Finally, a hybrid model, which combines a Random Forest with a Panel Smooth Transition model, is created. The results show that the Panel Smooth Transition Regression model is able to outperform the benchmark model and is better suited to deal with the abnormal data collected during the COVID-19 pandemic. The hybrid model is also able to outperform the benchmark model but returns similar results to the relatively easier Panel Smooth Transition Regression model.

Contents

1 Introduction

In early 2020, the fear of the new coronavirus spreading around the world, combined with an ongoing collapse of the oil price, resulted in a crash of 7% on the S&P500 and triggered an emergency halt in a few minutes after market opening. In just a few weeks time the S&P500 lost over $5 trillion dollar (Ballentine et al. (2020), Randewich (2020)).

This novel virus, which emerged in late 2019 and is known as COVID-19, spread rapidly around the globe. In just a few months the World Health Organisation declared the virus to be a global pandemic[1]. This abrupt and unexpected event brought changes to the economies worldwide.

The virus brought, and will bring even more, challenges for credit risk and corresponding credit risk models. As a result of the start of the COVID-19 pandemic, various companies in various industries were limited in their work or even brought to a halt. Companies were no longer able to sell their products or services due to measures that governments have taken to limit the movement of people and stop the spreading of the virus. Those companies came into a position with limited to no income, bringing the risk drivers to extreme values. Which in turn results in the credit risk models assigning higher credit scores to those companies due to the worsening of the economy during the pandemic.

At the same time, banks and governments are employing special measures in order to prevent the companies from going into default. These measures include, for example, support packages from the government in which companies can get special allowances or relaxed rules with respect to payment of taxes and special payment holidays from the banks. As a result of these measures, the relations that used to exist between the variables in the credit risk models, no longer behave the same.

Besides the possible incorrect functioning of the current credit risk models, the future credit risk models will suffer from the same issues. Credit risk models make use of historical data. Once the COVID-19 pandemic is over, the historical data at hand will still have a period that includes the COVID-19 pandemic which the models might not be able to properly handle. As a result of the earlier mentioned issues, the data collected during the pandemic might no longer be adequate to be used in the credit risk models in the coming years. Currently, the models will likely be adjusted by the bank based on expert opinion, but this is often not as desirable as when the model could be adjusted.

The full effects of the COVID-19 pandemic are yet to be seen, especially since the pandemic is not yet over. However, there is already data available since the start of the pandemic in The

[1] https://www.bbc.com/news/world-51235105

Netherlands. This data will give a good starting point to adjust the current credit risk models if necessary. The aim of this paper is to investigate whether the credit risk models do indeed perform worse during the COVID-19 pandemic. After which new models will be proposed and tested to see whether the aforementioned problem can be solved statistically instead of solving it based on expert opinion.

The severity of the impact on each sectors will most likely differ based on the sector the company is active in. Tourism, restaurants, and the oil industry are among the companies that will likely be highly affected by the pandemic while at the same time supermarkets and online stores will experience no impact or might even be growing Mirza et al. (2020). Therefore, during the research, it will be important to keep the different sectors in mind.

The paper starts by investigating whether there is indeed a change visible in the financial data collected in the year 2020, being the only year that the COVID-19 pandemic was taking place at the moment of writing, compared to the years before. Both visually and statistically, the data seems to confirm that the behaviour in 2020 is different compared to the earlier years.

A total of 4 different models are estimated. Each of the models is estimated using two different data sets. The first data set will be used to evaluate the performance of the model during the pandemic when the model is estimated only using the pre-pandemic data. The last data set will be used to estimate the final model using all the data. This method ensures that both the overall model performance as well as the ability to deal with COVID-19 pandemic data can be evaluated.

A Logistic Regression model is created as a benchmark model. This model will be used to compare the performance of the other models. Using the model that is estimated on only the pre-pandemic data, the performance of the benchmark model decreased during 2020 with roughly 48 percent more incorrect default predictions. Furthermore, the accuracy decreased from 76% to 29%. These results confirm the earlier mentioned problem that the financial variables changed in such a way that the model fails to capture the change.

The second model is the Panel Smooth Transition Regression model. The model adds each of the variables from the Logistic Regression model again but adjusted by a continuous transition variable. The transition variable adjusts the model to the state of the economy and should therefore improve the adjustability of the model to the actual state of the economy. The Panel Smooth Transition Regression model manages to keep its performance when applied to the data of the COVID-19 pandemic while not being calibrated using any of the COVID-19 pandemic data. The model has an accuracy of 77% and is able to identify 69% of all the default cases during the COVID-19 period suggesting that the Panel Smooth Transition Regression model is

better able to deal with the COVID-19 pandemic data than the benchmark model.

The final model is a hybrid model that aims to combine a Random Forest model with a Panel Smooth Transition Regression model. A Random Forest model is known to be able to outperform a Logistic Regression model but is not widely used since it is a rather difficult to interpret model which is undesirable for credit risk models. However, the hybrid model uses the results of the Random Forest model to find the optimal combination of variables and interaction variables to be used in the Panel Smooth Transition Regression model. This way, the hybrid model tries to use the feature selection power of the Random Forest model and combines it with the interpretability of the Panel Smooth Transition Regression model.

The paper starts, in section 2, with a literature review after which in section 3 the used data will be described and corresponding data cleaning methods. In section 4 and section 5 the methodology used will be described with the corresponding results respectively.

2 Literature

Since the COVID-19 pandemic only started recently, little literature has been published. Mirza et al. (2020) analysed the impact of the COVID-19 pandemic on the corporate solvency of non-financial listed companies in the European Union. They found that the probability of default increased for all firms as a result of a decrease in market capitalisation. Especially companies active in the manufacturing, mining, and retail sector are vulnerable. As a result of the lockdowns in many European countries, the revenue of many companies contracted. The European Union even experienced unprecedented economic costs.

For the credit risk models to be functional with the pandemic data, it likely requires changes to the current models. A Logistic Regression model is currently one of the most used models in the banking industry (Dong et al. (2012)). Koenker and Bassett (1978) noticed that this methodology uses the median value of the independent variable and thus might not be robust to outliers. They created an alternative model, called the Quantile Regression model, in which the model uses the whole conditional distribution instead of only the mean of the data. This method is more robust and flexible compared to the least square model. By splitting the data into various smaller groups the model is able to capture non-linear relations in the data that might be caused by unforeseen events(Kuan et al. (2013)).

Kuan et al. (2013) created a multiple threshold Quantile Regression model based on the aforementioned Quantile Regression model. They adapted the original model since the theory behind an economic model does not suggest the number of regime changes a priori, and thus making it difficult to create the required Quantiles. In the multiple threshold Quantile Regres-

sion, the Quantiles are created during the estimation of the model while minimising the error term.

However, since a crisis is almost always unique in severity, combined with the potentially limited amount of data available from earlier shorter crises, it will be difficult to estimate a Quantile Regression model. Furthermore, the estimation of the parameters and the quantiles will be computational expensive and time consuming and may therefore not be desirable.

Kadilli and Markov (2012) came with the Panel smooth Transition Regression (PSTR) modelling approach in which two opposite regimes can be modelled. The model essentially copies the variables of a Regression model and adds them to the regression model but adjusted to the state of the economy using a smooth transition parameter. The model can switch between two regimes via the use of a smooth transition parameter. The advantage of this model, compared to the Quantile Regression model, is that it is also able to capture an hypothetically unlimited amount of different regimes in between the two regimes due to the smooth transition parameter.

With the PSTR modelling approach, Kadilli and Markov (2012) tried to uncover the presence of non-linearity in the determinants of the inflation forecasts of the European Central Bank. They used this method to find non-linearity between individuals and over time. They estimated two regimes in which the first regime refers to a normal period and the second regime to a crisis period. The switches between the regimes are achieved with the use of an transition variable. The transition variable is a continuous transition function which is bounded between 0 and 1 and determines the switch between the two regimes. The continuity of the transition function allows for a theoretically infinite number of intermediate regimes which allows the model to capture various scenarios in between 2 extremes. The main advantage of this method is that it is a simple parametric approach to introduce non-linearity and is only a slight adjustment to a Logistic Regression model which currently is the market practice model for credit risk (Dong et al., 2012). Secondly, it allows for individual heterogeneity and time variability between the variables. The model also allows for varying marginal effects of the explanatory variables based on the state the economy is in. The downside of the PSTR model is that it might be difficult to identify the optimal observable economic indicator that is able to appropriately adjust the model to the state of the economy.

Machine learning techniques are known to be able to create more accurate models compared to a Logistic Regression model (Couronné et al., 2018). These techniques are currently not widely used in the banking industry due to the complexity and due to being difficult to interpret. However, some researches have been published that use these techniques on credit risk. Khandani et al. (2010) applied machine learning techniques to construct a non-linear non-parametric model

to predict credit card consumer credit risk. The technique used is a generalized Classification And Regression Trees (CART) model. They found that by use of this model, an out-of-sample forecast was achieved that significantly improved the classification rates of delinquencies and defaults. A CART model still creates relatively easy interpretable decision rules compared to other machine learning techniques which make this technique attractive.

The downside of a CART model is that it may lead to high variance estimates due to overfitting. A possible solution for this is to aggregate multiple CART models. When multiple CART models are aggregated, the variance of the estimates is reduced due to the decorrelation of the individual models. This combining of multiple CART models is a technique known as a Random Forest. A Random Forest generally leads to even better accuracy and improved probability estimates (Kruppa et al. (2013)).

However, a Random Forest does lead to reduced interpretability. Lundberg and Lee (2017) introduced a novel methodology to make the predictions of these complicated models interpretable. This is achieved by assigning relative importance values to all the variables in the model. The framework for this is called SHAP (SHapley Additive exPlanations). SHAP assigns an importance value to each variable by investigating what the marginal impact of including a certain variable in the model is. This returns a list with the importance value of each variable which in turn results in an easier to interpret model by showing which variable is the most important in the machine learning technique.

levy and James O'Malley (2020) proposed a hybrid methodology that aims to combine the strength of a machine learning model with the interpretability of simple Linear Regressions or Logistic Regression models. The methodology consists of two steps. The first step uses a machine learning technique to find the best variables which are then used, in the second step, in a regression model. Their idea is that if the Logistic Regression itself would be the true model, it would outperform any other method, but if the regression model would not be accurate it could be improved with an approach that uses powerful search capabilities. The methodology uses the aforementioned SHAP values to assign a relative importance value on the variables of the machine learning technique. The complex predictors that are then found by the machine learning technique are added to the Logistic Regression which would be able to bring the model closer to the true values than other competitor techniques. Their results showed that a hybrid model, that combines a Random Forest and a Logistic Regression model, is able to achieve performances similar to less interpretable approaches. The hybrid model was even able to achieve similar or better performance compared to the Random Forest approach on some data sets.

3 Data

The data that is used for this research is raw data provided Zanders. The data contains financial information of various companies from all over the world. This information is mostly related to the financial side of a company such as the operating revenue and gross profit. It also contains some other information such as the number of employees and the size of the company. In total, there are 63 variables present for each observation. A table with all the variables present can be found in appendix A. Each row contains information from a single company for a single year.

Due to the data collection method used to create the database, it is not possible to see the exact month the financial data are representing. For this reason, the date corresponding to each observation will be based on the year in which the data has been published

3.1 Data processing

The total data set has over 400 million rows containing information from companies all over the world. Since the main focus of the research will be on the Dutch Market, the first step will be to remove all companies that are not located within The Netherlands. Removing the data of other countries shrinks the data set to nearly 20 million observations for the Dutch companies. Each of these observations has 63 variables which contain information such as financial data, an identification number for the company, and a status indicator whether the company has defaulted.

3.1.1 Missing data

A relatively large part of the data set is missing. These missing data points need to be dealt with before they can be used in the models. The first step is to remove the observations that has no financial data available at all. If an observation has no financial data, it is of no use in the model and can thus be removed.

Secondly, each financial variable will be checked on the percentage of missing data. If a specific financial variable is missing more than 40% of its observations, the whole variable will be removed since it will be too difficult to appropriately fill in these missing variables (Jakobsen et al., 2017). This process will decrease the data set to nearly 8 million observations with 13 financial variables.

To not shrink the data set any further, the remaining missing data points will be replaced with the average value of the corresponding variable. This method is chosen since it is computationally inexpensive and ensures that the mean of the variables remains unchanged. The downside of this method is that it ignores the relationship between the other variables and underestimates

the variance (Zhang, 2016). Better results may be achieved using other imputation methods. However, this is beyond the scope of the current research which has the goal to improve the model and not to improve the data set.

3.1.2 Financial ratios

Some of the financial variables will be converted into financial ratios in line with the financial ratios used by Zanders which are based upon expert opinions. Financial institution often use financial ratios to determine the credit worthiness of its borrowers due to the useful information the ratios provide (Beaver, 1966). The advantage of using these financial ratios is that they are more informative about the financial health of a company compared to the financial variables itself. For example, if two companies have the same assets value but different liabilities, the company with the lowest liabilities is likely to be in a better financial situation.

The following financial ratios will be used:

$$Currentratio = \frac{Currentassets}{currentliabilities}$$

which is the liquidity ratio that measures whether the company has enough resources to pay its short term obligations.

$$TangibleNetworth = \frac{Totalassets - totalliabilities - intangibleassets}{1000}$$

which is a method used to calculate the net worth of a company excluding its intangible assets such as intellectual properties.

$$Solvency = \frac{Tangiblenetworth}{Totalassets}$$

which is the ability of a company to meets its long term financial obligations.

The financial variables used to calculate these ratios will be removed from the list of predictors since they will be highly correlated with the corresponding ratio. A total of 11 variables will remain after replacing the variables by the earlier mentioned ratios.

3.2 Multicollinearity

A problem that may arise during the estimation of a regression model is the presence of multicollinearity. Multicollineairity may arise when variables are highly correlated with each other. As a consequence of this correlation, the estimators of the coefficients will get large variances. High variances in turn may lead to estimates that themselves are higher or may have signs that are not in line with the theoretically expectations of those variables (Mansfield and Helms, 1982).

To detect the possible presence of multicollinearity the Variance Inflation Factor (VIF) can be used. The VIF is an indicators that shows how many times the variance of the coefficient will be inflated due to the presence of multicollinearity. In an optimal model, the VIF would be equal to 1 for every variable. Since the VIF is based on the relation of the variable with all the other variables, deleting one variable would change the VIF of all correlated variables. Therefore, deleting variables based on the VIF should be done in an iteratively manner. A VIF value of around 5 and higher is often considered to be highly correlated (Daoud, 2017). Furthermore, a correlation value of above 0.85 is already sufficiently high that it could lead to multicollinearity (Schroeder et al., 1990).

To prevent the risk of multicollinearity, the variables with the highest VIF will be removed iteratively until an acceptable VIF of around 5 is achieved. The VIF for each variable can be seen in table 1.

Table 1: The Variance Inflation Factors (VIF) for each variable in the data set. The virable with the highest value will iteratively be removed to prevent multicollinearity. Going from left to right, the variable with the highest Variance Inflation Factor is removed.

Variable	VIF (all variables)	VIF (Shareholders funds removed)
Stock	1.22	1.22
Debtors	1.45	1.45
Other current assets	6.02	5.97
Cash-cash equivalent	6.10	6.05
Shareholders funds	13.13	Removed
Capital	1.09	1.08
Other shareholders funds	2.74	2.43
Total shareholders funds liabilities	3.01	2.44
Current ratio	1.00	1.00
Tangible net worth	9.24	3.06
Solvency	1.00	1.00

The Shareholders funds variable had a VIF of just over 13. This means that it is highly correlated with at least some of the other variables. This is confirmed when Looking at the correlation values. The Shareholders funds variable has a correlation value of 0.94 with the Tangible Net Worth variable and 0.79 with Other shareholders fund which could therefore lead to multicollinearity. A table with all the correlation values can be found in appendix B.

After removing the Shareholders funds variable, almost all of the VIF decrease for the other

variables due to the correlation the Shareholders funds variable has with most other variables.

This time the Cash-cash equivalent variable has the highest VIF value of slight above 6. This variable is just above the accepted value of 5. However, looking at the correlation values, the highest correlation value for the Cash-cash equivalent variable is 0.81 and is thus not directly an indication for the possible presence of multicollinearity (Schroeder et al., 1990). To not decrease the data set any further it is decided to keep the variable in the data set and only remove the Shareholders funds variable.

3.3 Data exploration

As mentioned, the data set that will remain after the initial filtering contains nearly 8 million observations with each observation containing 10 financial variables. Besides the 10 financial variables each observation has a variable for the year, an indicator whether the company has defaulted that year, an unique ID number to differentiate the various companies, and a variable that indicates the industry the company is active in. The number of observations and defaults for each year are given in table 2.

Table 2: Total number of observations and defaults in each year that is present in the data set.

year	observations	defaults	year	observations	defaults	year	observations	defaults	year	observations	defaults
1984	12	0	1994	33818	0	2004	88081	6	2014	662204	652
1985	16	0	1995	39351	0	2005	102518	19	2015	672216	482
1986	16	0	1996	44420	0	2006	119183	7	2016	691398	392
1987	18	0	1997	43558	0	2007	127766	7	2017	684009	467
1988	33	0	1998	43824	0	2008	194423	13	2018	679812	403
1989	49	0	1999	47445	0	2009	331119	22	2019	589716	321
1990	54	0	2000	48608	1	2010	592388	29	2020	10335	72
1991	52	0	2001	59223	0	2011	610990	37	2021	1	0
1992	81	0	2002	77754	5	2012	624178	68			
1993	28800	0	2003	82916	1	2013	644155	374			

As is visible, the years 1984 until 1992 and 2021 are very limited in number of observations and have no defaults present. From the year 2000 onward there are at least 10000 observations available and at least one default each year, with the exception for 2001 with no defaults. To limit the noise of data that is too far in the past or of data that might not fully represent the actual period, the data that will be used is from 2000 until 2020. This means that the data set that will be used for the modelling has a total of nearly 7.7 million observation with just short of 3400 defaults.

Since the impact of the COVID-19 pandemic is likely to be different for each industry, it is of interest to take the various industries into account. Figure 1 has the distribution of the

industries over the selected years. The figure shows that the distribution of the various industries is not equal. The Financial, Wholesale, and Manufacturing industry are the largest sectors while the Mining, Water, and Electricity industries are the smallest. The model will thus likely be skewed in favor of the industries that are most present in the data. However, this should not be a problem since the distributions of the industries remain roughly the same throughout the years.

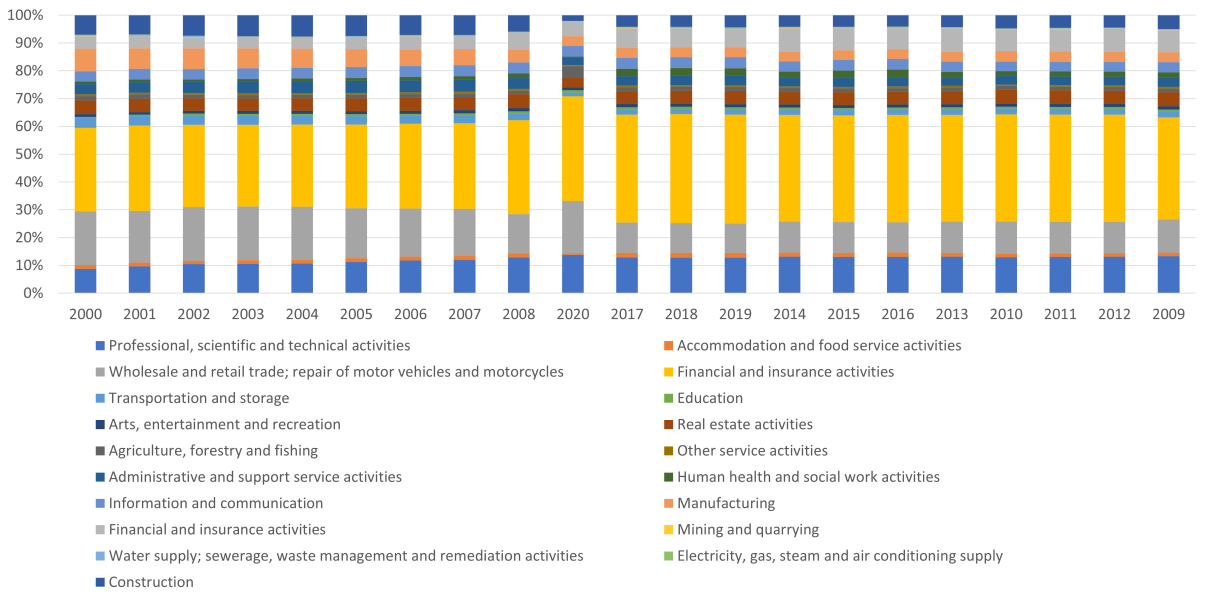

Figure 1: Distribution of the various industries present in each year.

3.4 Outliers

Before the data can be used, it is important to remove possible outliers. An outlier is an observation that has a certain value for its variables that is way higher or lower compared to the other observations. This can be due to an collection error or another mistake that took place during the processing of the data. Keeping outliers in the data set may overvalue the model and cause the estimate to vary drastically from the true model (Ghosh and Vogt, 2012).

Based on the descriptive statistics of the variables, it can be concluded that the data likely contains some outlier. The maximum and minimum value of the three most interesting variables are given in table 3. As can be seen in the table, there is at least one observation present in the data set with a Total shareholders fund value of slightly over 256 billion euros and an observation with a negative Other shareholders fund of 5 billion euro. These values are not realistic and are therefore likely the results of an error in the data collection method. The full table with the statistics for all the variables can be found in appendix C.

Table 3: Maximum and Minimum observed values of three variables in the data set with the outlier included and when the outliers are removed.

	With outlier		Removed outlier	
	Minimum	Maximum	Minimum	Maximum
Debtors	-2.42E+09	2.02487E+11	-330908000	461605193
Oher shareholders funds	-5.54E+11	1.51109E+11	-980222367	998864000
Total shareholders funds liabilities	-1.33E+10	2.56962E+11	-308807000	1706341805

A Z-score test will be used to check whether any of the observations can be considered as outliers. The Z-score test marks an observation as an outlier if the value is located at more than 3 standard deviations from the mean of the corresponding variable (Osborne, 2004). by use of the Z-score test, slightly more than 16000 observations are marked as outliers and will be removed from the data set.

Table 3 has again the maximum and minimum value of the same variables as before but this time after removing the outliers using the Z-score test. Whereas the maximum value for the Total shareholders fund was over 256 billion, the maximum is just shy of 2 billion after removing the outliers which is a more realistic value. The full descriptive table of the variables after removing the outliers can be found in appendix C.

The remaining variables will be standardised. This is done since the range between the various variables is quite large. Standardising the variables ensures that the range of the variables are comparable and that the estimated coefficients are easier to interpret on their relative impact on the model.

3.5 Balanced data

The current data set consist of 8 million non-default observations and 3400 defaults, which is roughly a ratio of 1 defaults for every 2350 non-defaults. If this ratio is not taken into account in the modeling phase, the model will likely set all predictions to non-default, giving the model an accuracy of nearly 100%. However, for credit risk modelling, such a model would be useless since it does not predict any defaults. For the model to be able to have a practical prediction, it is thus important that the number of non-default and default observations is roughly the same.

Various methods can be applied to artificially adjust the ratio between non-default and default observations to partially resolve the aforementioned problem. In section 3.5.1 and 3.5.2 a weight balancing and an oversampling method will be described. Besides these two methods, a naive approach will be tested as well which does not adjust for the unbalanced data. The

naive approach will be used to check whether the assumption is correct that the model would only predict non-defaults if the unbalanced data is not accounted for. The method that has the best performance will then be used for the remaining of the paper.

3.5.1 Weight balancing

The first method to counter the unbalanced data is to attach weights to each observations. The weights assigned to the observations are inversely proportional to the frequency of that observation. This means that the default observations will receive a respectively high weight in the modelling part compared to the non-default observations and thus will be of higher importance to be properly predicted. Equation 1 has the formula for this method.

$$w_i = \frac{n}{c * s_i} \qquad where\ i = 0, 1 \tag{1}$$

where w_i is the weight for class i (i being a 0 for non-default and 1 for default), n is the total number of observations, c is the number of classes, and s_i the number of observation corresponding to the class.

3.5.2 Categorical SMOTE

Categorical SMOTE is, as far as we are aware, a novel adjustment to an oversampling technique that can be used to correct the proportional distribution between the classes by artificially creating more observations for the minority class. Categorical SMOTE is an extension to an oversampling technique known as the Synthetic Minority Over-Sampling TEchnique (SMOTE). SMOTE is a technique proposed by Chawla et al. (2002) that artificially creates new data points for the minority class based on the existing data.

SMOTE chooses a data point, x_i, from the minority class at random and selects the k-nearest neighbours to that data point from the minority class as well. The value for k is set to 5 being the standard value used in this method following the implementatation of Chawla et al. (2002). From the selected k-nearest neighbours, one data point, $x_{i,k}$, will be selected at random. The new data point, x_{new} is then generated using the formula in equation 2.

$$x_{new} = x_i + \sigma(x_{i,k} - x_i) \tag{2}$$

Where σ is a random number between 0 and 1. Using this method a new data point is created at random via the use of interpolation.

Since SMOTE interpolates the data at random, it does not take the industry the company is active in and the year the observation has been collected in into account. Using SMOTE on

all the data would mean that the industry effects and the year effects are ignored and that the new data points would be generated based on all the available data.

To ensure that the individual effects remain in the data, the data will be pre-processed in the Categorical SMOTE method. This method groups the data based on the year and the industry the company is active in. Each of these subsamples then contains data with comparable individual effects. After grouping the observations, the earlier explained SMOTE technique can be applied on each of these subsamples to ensure that the individual effects remain present.

This methodology of subsampling might have a few issues on the data set. First of all, since the number of defaults in the whole data set is relatively small, it is likely that a subsample will not contain any default observations at all. In this case, it will not be possible to resample from that specific data set. The second issue is in case a subsample only contains a small number of default observations. Oversampling using SMOTE will then lead to a cluster of new observations around the existing few default observations. The characteristics of the default data then might no longer be representative since all the observations will share almost the same characteristics, which will skew the data. Therefore, the proposed technique will require a minimum of 6 defaults to be present in the subsample in order for that subsample to be oversampled using the Categorical SMOTE technique. The value of 6 is chosen since SMOTE selects a data point and the 5 nearest neighbours thus requiring at least 6 data points to be present.

After these steps, the number of observations in each class should be closer together. However, it will not yet be a $1:1$ ratio due to the skipped subsamples that did not meet the threshold number of defaults. The weighing technique, which was explained in section 3.5.1, will be used to bridge the remaining difference between the two classes.

3.6 Economic indicators

The Panel Smooth Transition Regression model requires an economic indicator. This indicator is used to adjust the model each year according to how the economy behaves in that year. Since there is no unambiguously answer to what the best economic indicator is, a multitude of indicators will be tested to see which gives the best results.

The indicators that will be used are based on varying events in the world. The indicators are: GBP, 'Consumer convidence', and 'willingness to buy' in The Netherlands [2]. Some of these indicators are available on a higher frequency (daily, weekly, etc.) than the data that is used

[2]The values of the economic indicators are retrieved from https://opendata.cbs.nl. CBS is a dutch governmental institution that collect statistical information about The Netherlands.

to calibrate the model (yearly). If this is the case, the value the indicator has on the 31st of December will be used for the corresponding year.

4 Methodology

Up until here, the assumption has been made that the data from 2020 onwards is indeed different compared to the years before. The first step is to see whether the data actually confirms this assumption before any models can be created. This will be tested using two different methods as will be explained in section 4.1. In case the data does not confirm the assumption, the models will not be estimated on the data since the assumed problem is likely not yet present in the data.

The available data consists of information for various companies spanning across multiple years. This type of data is referred to as Panel Data. Panel Data requires some additional attention since its focus is on multiple observations at each point in time, unlike regular time series analysis which focuses only on one observation at each point in time. Section 4.2 will explain the applied methodology to deal with Panel Data.

This paper consist of four different models that will be used. At first, in section 4.3, a Logistic Regression will be explained. The Logistic Regression model is defined as the benchmark model which is used as the baseline for the performance measurements of the alternative models, which aim to outperform the benchmark model. Then in section 4.4 and 4.5 the methodology for the Panel Smooth Transition Regression model and a Random Forest model will be explained. These alternative models aim to better capture the effects of the crisis as compared to the benchmark model. The last evaluated model is a hybrid model that aims to combine the power of the Random Forest model with the interpretability of the Panel Smooth Transition Regression model and will be explained in section 4.6. For these models to be considered effective, they should be able to outperform the benchmark model, which will be evaluated using the various the performance measures outlined in section 4.7.

4.1 Problem setting

The paper builds on the assumption that the COVID-19 pandemic had such an impact on the financial data that the credit risk models are no longer able to properly function from the year 2020 onwards. However, this is merely an assumption which should be verified before any novel models can be proposed.

To verify whether the behaviour of the financial data in 2020 is indeed different, the delta value for each variable will be calculated. The delta of a variable is the change in value between

two consecutive years and is calculated using the following formula:

$$\Delta_i, t = x_{i,t} - x_{i,t-1}$$

where $x_{i,t}$ is the financial variable i, and t is the period in which the observation is collected.

If the data did not have any big chances in 2020, the distribution of the delta between 2018 and 2019 should be roughly the same compared to the delta between 2019 and 2020. This is checked both visually and statistically. The delta between 2017 and 2018 is compared as well to ensure that the data behave consistently in the years before and that the change to 2020 is indeed unique.

At first, a histogram will be made for each variable. This way an initial conclusion can be made whether the distributions looks the same. After that, by use of a t-test, the hypothesis that is made based on the visual inspection can be verified statistically. A t-test is used to check whether the distribution of a sample is significantly different compared to the distribution of another sample. If the t-test returns a significant result, it can be concluded that the samples were extracted from two different distributions and thus that the change in data is significantly different in between the years (Kim, 2015). In case these tests confirm the assumption that there is a different change in at least some of the variable going into 2020 compared to the years before, the models can be estimated according to the methodology explained hereafter.

4.2 Panel Data

Since the data set that is used is multidimensional data, which is known as Panel Data, the modelling requires special attention. Ordinary time series regression uses data of one company that is collected at several points in time. Panel Data has, besides the observations that are collected at each point in time for one company, observations of multiple companies that are collected at each point in time. Applying time series methods to Panel Data would discard the individual behaviours of each company and would assume that each company behaves the same. If the behavioural effects of the companies are not comparable, not incorporating the individual behavioural effects into the model could lead to an incorrect model since quite some information is being ignored.

Modelling Panel Data can be done in a multitude of ways. The first method is know as Pooled Panel Data regression. This method assumes that there are no unique effects for the different companies and the modelling is done in the same way as a regression model (Schmidheiny, 2021). The regression model is given in equation 3.

$$y_{it} = \alpha + \beta x_{it} + \epsilon_{it} \tag{3}$$

Where y_{it} is the default indicator and x_{it} the independent variables with $i = 1, ...N$ as the indicator for each of the N individual effects and t the indicator for the year.

The second method assumes heterogeneous Panel Data which means that individual effects are present in the data and is called Fixed Effect Panel Data (Torres-Reyna, 2007). If this is the case, ordinary time series methods can no longer be applied due to the aforementioned problem. The model can be adjusted via the use of dummy variables to capture the individual effects of the various companies (Baltagi et al., 2008). The equation becomes as given in equation 4.

$$y_{it} = \alpha_j d_{ij} + \beta x_{it} + \epsilon_{it} \tag{4}$$

The difference compared to the equation for the Pooled Panel Data regression is that the constant α is replaced with an individual effect α_j. The individual effect, α_j, is a set of individual variables which is multiplied by the dummy vector $d_{i,j}$ which is a vector of ones with length j, in which j is the indicator whether company i is active in the corresponding industry. Using the second methodology has the advantage that the fixed differences each company has on the probability of default are controlled (Johnson, 1995). However, if no clear difference is present among the industries, the former method is preferred since unnecessarily adding variables would not improve the model performance.

Adding a dummy variable for each individual would lead to perfect multicollinearity. Perfect multicollinearity is when an independent variable is a perfect linear combination of other independent variables (Allen, 1997). This would be the case if a dummy variable is added for each individual since one dummy variable can be predicted based on the value of the other dummy variables. To prevent perfect multicollinearity from happening, one dummy variable should be omitted for the individual effect.

Since the total data set contains more than one million unique companies, adding a dummy variable for each company will result in a computationally expensive model. Furthermore, the model would then not be applicable to companies that were not present during the model estimation and the interpretation of the model will become difficult. To counter this issue, the companies will be grouped by their respective industries. The companies can be split into 19 unique industries (a full list of industries can be found in appendix D). As mentioned before, one dummy variable should be omitted to prevent perfect multicollinearity meaning that 18 dummy variables will be added to model the individual effects.

Whether the Pooled Panel Data or the Fixed Effect Panel Data method should be used is depending on whether the model contains homoscedasticity and no autocorrelation. If either of these assumptions are non present, the Fixed Effect Panel Data model is preferred since the individual effects are likely to be present (Torres-Reyna, 2007).

Homoskedasticity is the assumption that the residuals are distributed with equal variance. Homoskedasticity can be tested using the Breusch-Pagan test (Rad et al., 2013). This tests whether the residuals become more spread out for higher values of the fitted values.

The second assumption that will be tested is the assumption of no autocorrelation. This will be tested using the Durbin Watson test (Baltagi et al., 2008). The Durbin Watson test will return a value between 0 and 4 where a value of 2 means no autocorrelation, a value between 0 and 2 positive autocorrelation and a value between 2 and 4 negative autocorrelation.

If either of these assumptions are not present, the Fixed Effect Panel Data is the preferred model.

4.2.1 Time variable

Furthermore, it is important to test whether the model requires a time effect variable. This will be tested by estimating a model that includes a dummy variable for each year. The model will then be evaluated on the coefficients of the time variables and its significance.

Two different versions to capture the time effect will be tested. The first version adds a single dummy variable for each year. The equation is similar as before but with the addition of δ_t which is multiplied by the dummy vector $d_{2,it}$ as can be seen in equation 5. The variable $d2_{2,t}$ is a matrix with ones for the corresponding year t and zeros otherwise and is multiplied by the estimated vector δ_t.

$$y_{it} = \alpha_i d_{1,ij} + \delta_t d_{2,t} + \beta x_{it} + \epsilon_{it} \tag{5}$$

The second method adds, instead of a dummy variable for each year, one single variable which contains the corresponding years as value. The equation then becomes as given in equation 6 with $d_{2,t}$ as the dummy variable that indicates the corresponding year and γ the estimated time effect.

$$y_{it} = \alpha_i d_{1,j} + \gamma d_{2,t} + \beta x_{it} + \epsilon_{it} \tag{6}$$

Both methods assume that unexpected changes or special events have happened throughout the years that the model itself is not fully able to capture. The difference, however, is that the first model assumes that the events throughout the years are different from each other and that each year has a unique impact on the model. The second method assumes that the time effect is constant throughout the years and that it could be an upward or downward trend.

Both models will be estimated and the time effect variables will be tested for its significance. In case the time variable is not significant, a time effect is apparently not present and should

not be added and the research will continue using equation 4.

4.2.2 Regularisation

Ridge Regularisation will be applied to the model during the estimation phase. Ridge regularisation is a shrinkage method used to reduce the estimated coefficients in the regression. Regularisation is applied in order mitigate the effect of mutlicollinearity. Multicollinearity is the effect that is present when one or more of the predictive variables are highly correlated to each other which is turn would lead to higher variance in the model and risks overfitting. Ridge regularisation works by adding a penalty term, $\lambda(\beta'\beta)$, to the loss function of the regression model. This penalty term punishes large coefficients and prevents overfitting by shrinking large coefficients to zero (Hoerl and Kennard, 1970). The optimal value for λ will be estimated using 10% of the data.

The various values for λ will be tested using a grid search cross validation. The scoring for the grid search will be based on the F1 score. The F1 score is a performance measure for the accuracy of the model and is calculated using equation 7.

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{7}$$

where

$$precision = \frac{TruePositive}{FalsePositive + TruePositive}$$

and

$$recall = \frac{TruePositive}{Truepositive + FalseNegative}.$$

. The exact calculation of the TruePositive, FalsePositive, and FalsePositive will be explained in section 4.7.

The F1 score is selected as scoring method since it puts the focus on the minority class in an imbalanced data set. Estimating an imbalanced data set would likely be able to reach an accuracy of nearly 100% by setting all observations to non-default, as explained in section 3.5, and is therefore inappropriate as scoring method.

The F1 score is a weighted average between the precision and recall values. Where precision is a metric that indicates the percentage of correct default predictions made and recall is a metric that indicates the percentage of correct default predictions made out of all possible default predictions (Jeni et al., 2013).

4.3 Benchmark model

The benchmark model is based on a Logistic Regression model. This choice is made because a Logistic Regression model is among the most used models in the banking industry due to its robustness and transparency. Setting an often used model as the benchmark model ensure that the results of the research are relevant (Dong et al. (2012)). The benchmark model uses the formula in equation 8.

$$Y = \frac{exp(X\beta)}{1 + exp(X\beta)} \tag{8}$$

Where Y is a binary vector with $y_i = 1$ if company i defaulted and $y_i = 0$ otherwise. The risk drivers are in the vector X, and β is the vector of estimated parameters for each risk driver.

As mentioned in section 4.2, the type of data that is used is know as Panel Data. Adding the individual effect dummies that the Fixed Effect Panel Data brings to the benchmark model results in a adjustment to the formula, as can be seen in equation 9.

$$PD = \frac{exp(D_1\alpha + X\beta)}{1 + exp(D_1\alpha + X\beta)} \tag{9}$$

Where the variable, α, is added to equation 8 which results in the new formula. With D_1 as dummy matrix indicating in which industry the company is active in with the corresponding estimated coefficients in vector α.

4.4 Panel Smooth Transition Regression

As first alternative model, a Panel Smooth Transition Regression (PSTR) model will be tested as possible improvement over the benchmark model. The PSTR model is a good alternative since it is still a relatively easy parametric approach that allows for non-linearity in the model. At the same time, the model can create a smooth transition between two extremes scenarios, which in this case could be a non-crisis and a crisis period. The main advantage of this model is that the smooth transition allows the model to appropriately follow the economy when it slowly changes between a crisis and a non crisis period and vice versa. In Section 4.4.1 the working of model and the variables will be described after which the estimation method for the PSTR model will be described in section 4.4.2.

4.4.1 PSTR model

The PSTR model uses the methodology proposed by González et al. (2017) and can be seen as an extension to a regression model. The methodology essentially combines a regression model twice with different estimated parameters. By incorporating a continuous transition parameter,

the model can gradually move from the first regime to the second regime. The PSTR regression model is given in equation 10.

$$y_{it} = \mu_i + \lambda_t + \beta_0' x_{it} + \beta_1' x_{it} g(q_{it}; \gamma, c) + u_{it} \tag{10}$$

Where y_{it} is the dependent variable for $i = 1, ..., N$ and $t = 1, ..., T$ with N as the cross-sectional dimension, being the different economic industries, of the data and T the time dimension. The NxT dimensional vector x_{it} contains the independent variables and u_{it} is the error term with dimension NxT. The parameters μ_i and λ_t indicate the individual effects corresponding to each industry and the time effects in the data. The values of β_0 and β_1 are estimated in the model and indicate the different regimes. The values of β_0 will be the values in the base case in which the transition function is equal to zero. In case the transition function is equal to zero, the PSTR model is reduced to a ordinary regression model.

Once the model start switching regimes, the values of β_1 will slowly be added via the transition function until they are completely added and the model is switched to the other regime. The values of β_1 can therefore be seen as the delta value between the two opposing economic regimes.

The continuous transition function, $g(q_t, \gamma, c)$, is bounded between 0 and 1. The transition function is a logistic distribution with a cumulative distribution function which is given in equation 11.

$$g(q_t, \gamma, c) = \frac{1}{1 + exp(-\gamma(q_t - c))}, \gamma > 0 \tag{11}$$

In the transition function, the coefficient c is the location parameter, and q_t is the transition parameter. The transition parameter q_t is an observable variable that will be used as input data and need to be selected a priori. The location parameter, c, is a threshold value that sets the transition function to exactly 0.5 if the observable transition parameter, q_t, reaches that value. It can be seen as the point at which the transition function is exactly between the two regime.

The parameters γ is used for the smoothness of the transition. The higher the value of γ the sharper the shape of the transition function will be and thus the quicker the model changes between the scenarios. The transition function becomes an instant switch between the 2 extreme scenarios as $\gamma \to \infty$. The parameters γ is assumed to be positive resulting in a strictly increasing transition function (Hurn et al., 2014).

The difficulty in the use of this model is that the observable transition variable, q_t, is not subjected to any restrictions by the model. The observable transition variable should be selected before the model can be estimated such that the variable is a good indicator for the state of the economy. If the selected observable transition variable fails to properly track the economy, it will likely also fail to adjust the PSTR model appropriately.

Another issue with this model is that it could be computationally difficult to get a precise estimate for the value of γ when the value of γ becomes large. If the value of γ is of a relatively different magnitude compared to the other parameters, the convergence will be slowed down. Getting a precise estimate for γ then requires many estimations with only small differences. Furthermore, the log-likelihood becomes rather flat when the parameter γ becomes large which could make the convergence even slower (González et al., 2017). In case the estimated value for γ becomes large, Goodwin et al. (2011) suggest the use of the following transformation to solve the problem:

$$\gamma = exp(\eta)$$

Using this transformation, the restriction for γ being strictly positive is removed. Furthermore, due to the shape of the transformation, the search focuses on a smaller range of parameters for η while keeping the original shape of γ (Hurn et al., 2014).

4.4.2 PSTR estimation

The estimation process of the PSTR model is done in two parts. At first, the values γ and c for the transition function will be estimated. This estimation will be done by using a grid search on a subsample of the data set of roughly 10% of the observations. The grid will be made of various values ranging from -5 to 5 for c and 0 to 10 for γ. A table with each combination from the grid search with some performance measures is located in appendix E. Each of the combinations in the grid search will be used to estimate the model with as target to minimise the sum of squared errors of the PSTR model as given in equation 12.

$$Q^c(\gamma, c) = \sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{y}_{it} - \hat{\beta}(\gamma, c)'\tilde{x}_{it}(\gamma, c))^2 \tag{12}$$

Where γ and c are the hyperparameters that are tested. The parameter $\hat{\beta}(\gamma, c)$ is obtained by estimating equation 10 using the selected hyperparameters and $\tilde{y}_{it}$ and $\tilde{x}_{it}$ are the default data and explanatory variables.

After estimating each combination in the grid seach, the parameters that give the lowest sum of squared errors will be selected for the estimation of the final PSTR model. In case the optimal combination of hyperparameters are border values, being the lowest or highest values in the chosen range of γ or c in the grid search, the range of values will be extended to ensure that the optimal combination is found.

In the second step, the transition function will be calculated using the optimal values for γ and c which were found in the previous step. After which the final PSTR model, in equation

10, will be estimated using the remaining data.

The starting values for γ and c that were selected in the previous step are used to estimate the other parameters conditional on the values of γ and c in equation 10.

4.5 Random Forest

The third model that is created is a Random Forest model. A Random Forest works by combining multiple independent Decision Trees. These Decision Trees use a randomly selected subset of variables to get to their own independent results (Belgiu and Drăguţ, 2016). The Random Forest combines the prediction of each Decision Tree and takes the average of the outcome of all these independent predictions to get to a final prediction. Since Random Forest models are known to be able to outperform Logistic Regression models (Couronné et al., 2018) (Pereira et al., 2018), it is an interesting next model to investigate. The additional randomness that is added by combining multiple independent Decision Trees protects the overall model against the individual errors of each tree. Furthermore, by use of the law of large numbers, the model is able to prevent overfitting and deal appropriately with outliers (Breiman, 2001) (Pal, 2007).

A Decision Tree works by making a split at each node in the tree based on the impurity. The impurity is based on the number of data points that are labelled incorrectly (Myles et al., 2004). The data that is used to make these splits is randomly selected but still requires some prior settings such as the depth of each tree. Furthermore the Random Forest itself also requires some prior settings such as the number of trees the model considers. These values will be determined by use of a grid search. The grid search will be applied on a range of possible values for these parameters and will be optimized using cross validation search. This method will evaluate various combinations of parameters and will return the optimal set of parameters that gives the best model performance.

The major downside of the Random Forest model is that it results in a black box in which it is not possible to fully see the process from beginning to the end. To get the full interpretation out of the Random Forest model, it requires an investigation into each separate decision tree that is created on the randomly selected data points. Even interpreting a single decision tree could be difficult if, for example, the depth of the tree is large.

4.6 Hybrid model

The final model is a model that combines the Random Forest model with the PSTR model. As mentioned before, the downside of the Random Forest is that it leads to a black box which makes the Random Forest less applicable. To counter this issue, a new methodology is used that

converts the Random Forest model into a hybrid two-step model following the methodology of levy and James O'Malley (2020). In the first step, a Random Forest will be created in line with the methodology explained in section 4.5. Each variable that is used in the Random Forest model will then be assigned a SHapley Additive exPlanation (SHAP) value following the method suggested by Lundberg and Lee (2017).

The SHAP value is designed to interpret a model and evaluate why the model makes certain decisions by assigning an importance value to each variable. The variable represents the impact of including that variable in the model, as will be explained in section 4.6.1. Not only does it look at the individual variables but also at the interaction terms that might be present in the model. An interaction term is a multiplication of two variables which might contain information that is not yet captured by the model if the variables are merely added separately. By running SHAP on the Random Forest model, it is possible to get a list of the most important individual variables and the combination of variables that have the highest impact in the Random Forest model.

In the second step, the selected variables and the combination of interacting variables with the highest importance values are added in the PSTR model. This methodology is chosen since it combines the power of the machine learning technique to reduce model risk, improve the model predictive power and variable selection (Couronné et al., 2018) with the interpretability of the PSTR model.

4.6.1 SHAP

Interpretability is among the most important requirements for credit risk models. Therefore, SHAP values will be used to open up the black box that the Random Forest created. SHAP values was initially created for game theory but was thereafter adapted and used to interpret various machine learning models (Fryer et al., 2021). The SHAP importance value is based on the marginal contribution a variable has to the model. The formula to calculate the importance value is given in equation 13.

$$\phi_i = \sum_{S \subseteq F \setminus i} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{s \cup i}(x_{S \cup i}) - f_S(x_S)] \tag{13}$$

Where ϕ_i is the SHAP value of variable i, F is the set of all explanatory variables, and S a subset of F. The effect is calculated with a model that includes the variable i, $(f_{S \cup i})$, and a model that excludes the variable i, (f_S). While x_S represents the values of the input variables. The absolute values in the formula are taken since the variables with the highest importance value should be selected regardless of it positive or negative.

Eventually, a list will be returned that indicate the variables that have the highest importance value in the model.

Since equation 13 might not be directly self evident of its working, the SHAP value can be explained by use of a set of axioms (Fryer et al., 2021). For explanational purposes we set the function model using variable set X to $G(X)$. Using this notation the following axioms can be imposed on ϕ:

Axiom 1. Efficiency: Combining the SHAP value for each variable i results in the full model.

$$\sum_{i \in F} \phi_i = G(F).$$

Axiom 2. dummy variable: if variable i is a dummy variable and contributes nothing to the model, then its SHAP value is 0.

$$[\forall S : G(S \cup i) = G(S)] \implies \phi_i = 0.$$

Axiom 3. symmetry: If the contribution of two features is equal then the SHAP values are equal as well:

$$[\forall S \setminus i, j : G(S \cup i) = G(S \cup j)] \implies \phi_i = \phi_j$$

The SHAP value can further be split into an interaction effect and a main effect for each variable. The interaction effect is when two variables combined result in a even higher impact on the model than the sum of the two individual variables effect. Interaction effect might be present in case the effect of variable i is dependent on the effect of j. By doing this, additional insight in gained into what exactly determines the outcome of the model. For the current paper, only pairwise interactions will be considered which eventually will lead to a matrix of pairwise interaction values. The calculations for the SHAP interaction value are done in a similar way as the SHAP value itself which was explained above. The SHAP interaction value are calculated using equation:

$$\phi_{i,j} = \sum_{S \subseteq F \setminus i,j} \frac{|S|!(|F| - |S| - 2)!}{2|F-1|!} \bigtriangledown_{i,j}(S) \quad \text{where } i \neq j \tag{14}$$

$$\bigtriangledown_{i,j}(S) = f_{s \cup (i,j)}(x_{S \cup (i,j)}) - f_{s \cup (i)}(x_{S \cup (i)}) - f_{s \cup (j)}(x_{S \cup (j)}) + f_{S(x_S)} \tag{15}$$

The SHAP interaction value is equally distributed between the feature i and feature j meaning that $\phi_{i,j} = \phi_{j,i}$. By combining equation 14 with equation 13, the individual effect for feature i can be calculated according to equation 16.

$$\phi_{ii} = \phi_i - \sum_{j \neq i} \phi_{i,j} \tag{16}$$

Using the SHAP values for feature selection from a machine learning model is an intuitively logical method. It extracts information from the model and then allocate a certain value to each variable indicating their impact in the model (Fryer et al., 2021). The downside of using the SHAP values is that the computational time increases exponentially (Aas et al., 2019).

4.7 Performance measures

The performance of the models is compared in two steps. The first step is to see whether the novel models can outperform the benchmark Logistic Regression model when tested on the data collected during the COVID-19 pandemic. In the second step, the models are estimated on the whole data set and tested whether they still can outperform the benchmark model when the COVID-19 pandemic data is only a small part of the data set. Theoretically, both the PSTR model and the hybrid model should be able to outperform the Logistic Regression model since the PSTR is an extension to the Logistic Regression model by adding more variables of which the coefficients can be equal to zero in case they do not have any predictive power.

The performance of the models will be evaluated by use of multiple performance measures. Following the paper of Lessmann et al. (2015), the models will be evaluated on the discriminatory ability using the Area Under ROC Curve (AUC), the accuracy of the probabilistic predictions (brier score), and on the correctness of the categorical predictions (classification error). Furthermore, the accuracy will be tested as the total percentage that the model correctly predicted the default status.

The results of a classification model can be split up into four groups according to a confusion matrix following the image in figure 2. The four possible outcomes are True Negative (TN) which are the correctly classified non-defaults, False Positive (FP) which are the non-defaults which are classified by the model as default, False Negative (FN) which are non-default classifications which are actually defaults, and True Positive (TP) which are correctly classified defaults. Since the default scenarios are of the most interesting cases, the percentage of False Negatives and True Positives will also be compared.

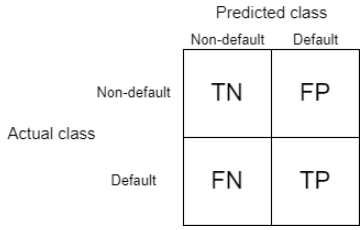

Figure 2: Four possible outcomes of a classification model. Where True Negative (TN) is the correctly classified defaults, False Negative (FN) being the defaults which are classified as non-default, False Positive (FP) is the non-defaults classified as default, and True Positive (TP) is the correctly classified defaults.

Besides the performance measures, the interpretability of the models will also be taken into account. Since it remains important for the model to be interpretable, a slightly higher accuracy might not be worth it if it results in a model that is harder to interpret.

The AUC measure represents the degree of separability in a classification model. The higher the value of this measure the better the model is capable of distinguishing between default and non-defaults cases. AUC is used as a summary of the Receiver Operator Characteristic (ROC) curve. The ROC curve plots the True Negative predictions against the False Positive predictions. In case the model is not able to distinguish between defaults and non-defaults, the ROC line will be a diagonal line and the AUC measure will get a value of 0.5. But the line has a sharp 90 degree corner in case the model can correctly dinstinguish between the defaults and non-defaults. The AUC is then a measure that reflects the area underneath the ROC, and thus a higher value indicates a better performing model.

The Brier score is used to measure the accuracy of the predictions with a value of 0 in case of total accuracy and a value of 1 if the model is completely inaccurate. The formula to calculate the brier score is given in equation 17.

$$Brier = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2 \tag{17}$$

Where N is the total number of observations, p_i is the predicted probability (being a value between 0 and 1) for observation i and o_i is the true outcome for observation i. A classification model returns a probability for what the outcome will be. The model then classifies the observation as either default or non-default depending on what has the highest probability. These probabilities are used in the calculation for the brier score to evaluate the accuracy of the model. In case the model is able to correctly predict all the observations, the brier score will return a value of 0. The model with the lowest brier score will be preferred over the other models.

5 Results

This section describes the results following the methodology as explained in section 4. At first, in section 5.1, the assumption that as a result of the COVID-19 pandemic a change has taken place in the data will be verified both visually and statistically. The use of the Fixed Effect Panel Data model or the Pooled Panel Data model will be tested in section 5.2. In section 5.3 the benchmark model will be estimated. After that, the PSTR model and the hybrid model are estimated in section 5.4 and 5.5 respectively.

Each of the models will be estimated twice. In the first version, the data for 2020 will be set aside and the model will be estimated using the remaining data. By doing this, the models can be verified whether they can properly estimate the defaults in 2020, being the COVID-19 pandemic period, using a model that was estimated without the data for 2020. The second model will be the complete model in which all the data will be used for the estimation.

5.1 Problem setting

Before any novel models can be created, the presence of the assumed problem, that the data behaves unexpectedly from 2020 onward, should be verified. In case the data does not confirm the assumption, the models can not be estimated on the data since the problem it should solve is not present in the data. This section consists of 2 subsections. In section 5.1.1 the data will be analysed visually. If the data visually seems to behave differently in 2020, the assumptions will be checked statistically in section 5.1.2.

5.1.1 Visual exploration

At first, the data will visually be checked to see whether the problem is present. This will be done by comparing the change in data, known as the delta, for each variable. The delta is calculated for each variable for the years 2017 to 2018, 2018 to 2019, and 2019 to 2020. By use of a histogram, it is relatively easy to see whether any big chances are present in the distribution of the delta value sets. In case no big changes have taken place, a comparatively similar distribution should be present and no big difference should be observable in the histograms.

The histograms of the variables that seems to have the most interesting differences are in figure 3 to 6. The histograms of the other variables are located in appendix F. In each of the figures, a histogram with the distribution of the Delta for 2017-2018, 2018-2019, and 2019-2020 are present.

Figure 3: Histogram for the delta values for the data between 2017-2018, 2018-2019, and 2019-2020 of the variable Intangible fixed assets. The y-axis indicates the fractions, being between 0 and 1, of the corresponding variable.

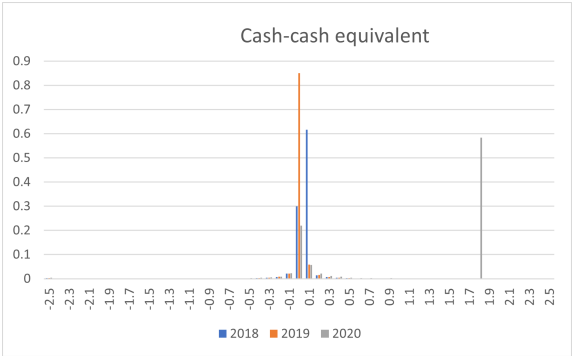

Figure 4: Histogram for the delta values for the data between 2017-2018, 2018-2019, and 2019-2020 of the variable Cash-cash equivalent. The y-axis indicates the fractions, being between 0 and 1, of the corresponding variable.

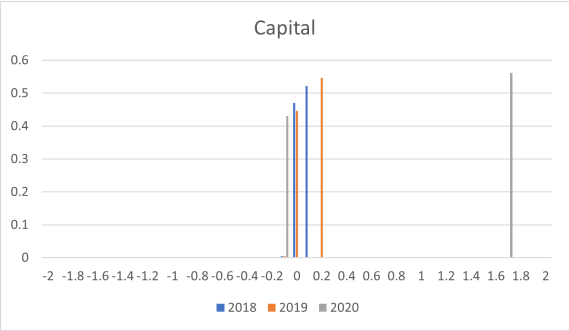

Figure 5: Histogram for the delta values for the data between 2017-2018, 2018-2019, and 2019-2020 of the variable Capital. The y-axis indicates the fractions, being between 0 and 1, of the corresponding variable.

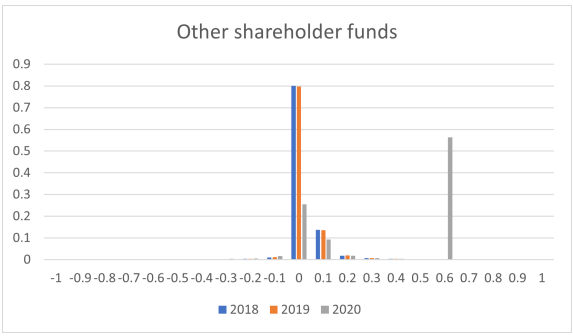

Figure 6: Histogram for the delta values for the data between 2017-2018, 2018-2019, and 2019-2020 of the variable Other shareholders funds. The y-axis indicates the fractions, being between 0 and 1, of the corresponding variable.

The delta distribution for the year 2019 to 2020 of each of the variables seems to have a peak on the higher end of the axis that was not present in the years before. The remaining of the histogram seem to have the bars relatively close to each other and often of similar magnitude. This means that between 2019 and 2020 the values changed in such a way that the difference between the two is larger compared to the change that took place in the years before. This leads to a preliminary conclusion that, for at least four variables, something did change in 2020 which was not present in the year before as can be seen in the figures.

5.1.2 Statistically exploration

To confirm the hypothesis that the delta did indeed change significantly between 2018 to 2019 and 2019 to 2020, a t-test is conducted on the same delta samples. A t-test is used to check

Table 4: The test statistics and corresponding p values of the t-test on the data sets with the delta value of 2018 to 2019 and 2019 to 2020. The variables that return a significant result at a 0.1% significance level are given in bold.

Variable	p value	Statistic	Variable	p value	Statistic
FixedAssets	0.356	-0.923	CashCashEquivalent	**0.000**	**-40.306**
IntangibleFixedAssets	**0.000**	**-12.720**	TotalAssets	0.589	-0.540
TangibleFixedAssets	**0.000**	**-5.339**	ShareholdersFunds	0.908	0.115
OtherFixedAssets	0.001	3.265	Capital	**0.000**	**-36.627**
CurrentAssets	0.411	0.822	OherShareholdersFunds	**0.000**	**-8.380**
Stock	0.332	0.970	NoncurrentLiabilities	0.004	-2.894
Debtors	0.017	2.380	CurrentLiabilities	0.194	1.298
OtherCurrentAssets	0.001	-3.277	TotalSharehFundsLiab	0.589	-0.540

whether the average difference between two samples is significantly different or that it can be explained by random chance. The results of the t-test are in table 4. The variables that are significant at 0.1% confidence interval are in bold.

The results confirm the earlier hypothesis that the four earlier mentioned variables are significantly different in 2020. Besides those four, one additional variable is significantly different at a 0.1% confidence interval suggesting that even more variables have changed in 2020 then visually expected. Based on this, it can be concluded that there have been a substantial change in values in 2020 possibly as a result of the COVID-19 pandemic.

5.2 Panel Data

At first, it is important to identify what kind of Panel Data the used data set is. As explained in the methodology in section 4.2, the modeling methodology is depending on whether the data has auto-correlation or heterogeneity. In case the model has no heterogeneity and no autocorrelation the Pooled Panel Data model can be used. If this is not the case, the Fixed Effect Panel Data model will be used. This will be tested using both the Durbin-Watson and Breusch-Pagan test respectively. The Durbin Watson test is a test that detects the presence of autocorrelation at a lag of 1 and the Breusch-Pagan test is used to detect heterogeneity.

The Durbin-Watson test detects the presence of autocorrelation at a lag of 1. The Null hypothesis is that there is no first order autocorrelation present and returns a value between 0 and 4. A value of 2 means no autocorrelation while a value below 2 means positive autocorrelation and a value above 2 is negative autocorrelation. Estimating the model on the initial 10% of the

data returns a Durbin-Watson test statistic of 0.063 meaning that the test suggest that there is positive autocorrelation present thus that the Pooled Panel Data model might not be the correct option.

To confirm this, the Breusch-Pagan test will still be used. The null hypothesis of the Breusch-Pagan test is that homoscedasticity is present in the data. Applying the test on the same model as before returns a p value of 0.000 meaning that the null hypothesis is rejected and heteroscedasticity is assumed to be present. The results of both these test appear to be reject the Pooled Panel model and thus the Fixed Effect Panel Data model will be used for the remaining of this research.

5.2.1 Time variable

The second step is to decide whether a time effect variable should be added and if so, what kind of variable. Two variation will be tested to see whether the time effect is present, following the methodology of section 4.2.1. The first model will contain a dummy variable for each year. This method assumes that certain unexpected events happened throughout the years which are not comparable in magnitude to each other. The second model will contain one variable with all the years. This methodology assumes that there is a certain trend going throughout the year. Whether a time variable should be added is depending on the significance of the time effect variable. If the time variable is not significant in either of the versions, there is apparently no time effect present and it should not be taken into account.

The first model that will be estimated is the model with a separate dummy variable for each year. If the estimated time dummies are significant in this model, it would suggest that certain unexpected events have happened in the years that the model was not able to capture and adding the time dummies would therefore improve the model.

After estimating the model, a p value of nearly 1 is returned for each dummy variable. This means that adding the dummy variables has no significant effect on the results of the model and could therefore be left out and thus rejecting the model with seperate time dummy variables.

The second model, with only 1 variable, for the time effect will still be tested. if this test returns a significant time variable, it would suggest that there is a time effect present which is constantly decreasing or increasing throughout the years instead of unique events each year. After estimating this model, a p value is returned of nearly 0 meaning that the time effect is significantly different from 0 and thus should be included in the model.

This concludes that the model that should be used in the research should include a single variable with the value for each year to capture the time effect. A possible reason why the single

time variable is significant while the time dummy variable were all insignificant could be due to a rising trend that is present in the data. This effect is apparently big enough that a variable should be added to capture this effect since the existing models were not able to do so.

5.3 Benchmark model

Since the data has shown to exhibit a change in data in 2020 compared to the years before, the next step is to start modelling the benchmark model. The benchmark model will be created using a Logistic Regression model as mentioned in section 4.3.

Before the benchmark model can be estimated, the hyperparameter ,λ, for the Ridge regularisation must be estimated, as explained in section 4.2.2, using 10% of the data. The remaining data will be randomly split into a set containing 80% of the data for the model estimation and 20% of the data will be used as an out-of-sample data set to test the performance of the model. In section 5.3.2, the results of the three different methods to deal with unbalanced data will be compared. The model that gives the best performance will then be used as the benchmark model and the balancing method used for the optimal model will then also be used in the further model estimations. In section 5.3.3 the Logistic Regression model will be used to proof the decrease in performance of the model when the model is used to predict defaults in 2020.

5.3.1 Regularisation

As mentioned in section 4.2.2 the model uses Ridge regularisation. Ridge regularisation shrinks the estimated coefficients towards zero to prevent overfitting. Various values for λ are tested using cross validation and compared for the optimal value for which the results can be found in table 5.

Table 5: F1 scoring results of the various parameter values for the Ridge regularisation. The data is estimated on 10% of the total available data.

λ	F1 score	λ	F1 score
1000	0.0009	1	0.0002
100	0.0006	0.5	0.0006
10	0.0008	0.2	0.0008
2	0.0009		

The model has been estimated using each of the values for λ. The data used for each estimation is based on the 10% of the data set apart for the hyperparameter tuning. The optimal value is selected based on the model that returns the highest F1 score. Based on these

results in table 5, an optimal λ of 2 is selected as optimal value and will be used for the model estimation.

5.3.2 Balancing method

Since the data is highly unbalanced with roughly 1 default for every 2350 non-default observations, it is important to choose the correct method to balance the data. This section will test three different balancing method as explained in section 3.5.

The first method is a relatively simple weighting method and is computationally inexpensive. A specific weight is assigned to both the default and the non-default classes to balance the data. A default case will receive a high weight while the observations with non-default will receive a low weight. This way, the model is punished relatively heavy in case it fails to identify a default case.

The second method is based on the Categorical SMOTE method as explained in section 3.5.2. This method combines an adjusted version of the SMOTE oversampling technique, that artificially creates new data points for the minority class, with the earlier mentioned weighting method. This method requires more work and is computationally more expensive compared to only using the weighting method. However, if this results in a higher model accuracy and better performance, the additional computational time could be worth it.

The last method that will be tested is the naive balancing method. This method takes each observation as it is and does not correct for the unbalancedness in the data. This method will likely results in a high accuracy since it will identify each observation as non-default. However, the default cases are the main interest of this paper and thus the naive method would not be selected since it fails to predict any defaults.

The training data set is based on 80% of the data while the remaining 20% is set aside for out-of-sample testing. The in-sample performance measures for each of the three methods are in table 6 with the corresponding confusion matrices in table 7.

Table 6: In-sample model performance for the benchmark model using three different methods to deal with the unbalanced data. The naive model does not account for the unbalanced data, the weighting method assigns a relatively high weight to the minority class and the Categorical SMOTE model combines an adjusted SMOTE oversampling method with the weighting method.

	naive	weighting	Categorical SMOTE
ROC AUC	0.80	0.81	0.74
Brier score	0.00	0.20	0.20
correct positive	0.00%	73.81%	72.23%
false-positive	0.00%	27.49%	23.44%
accuracy	99.96%	72.50%	65.90%

Table 7: Confusion matrix for the three methods to deal with unbalanced data. Going from left to right is the matrix for the naive method, the weighting method, and the Categorical SMOTE method. Each method is based on the same data set, but since the Categorical SMOTE method artificially creates default cases, the total number of in-sample observations is higher.

	0	1
0	5523934	5
1	2451	0

	0	1
0	4004565	1519374
1	642	1809

	0	1
0	3811449	2340247
0	1064181	2767357

The naive method seems to outperform the other two methods based on accuracy and the Brier score. This is as expected since the model almost only predicts non-default cases, as can be seen in the confusion matrix in table 7, which results in an accuracy of nearly 100% due to the unbalanced data. However, since the paper is mainly interested in the correctly predicted default cases, the naive method fails with 0% correctly identified defaults.

The weighting method seems to be the best option if compared with the categorical SMOTE method. The Categorical SMOTE method is in-sample only able to get a slightly lower percentage of false-positive prediction but is outperformed on every other measures suggesting that the weighting method is the optimal option based on the in-sample data.

Since the Categorical SMOTE method artificially creates new default observations for the training data set, as can be seen in the corresponding confusion matrices, a comparison between the performance measures based on the training data will not be a completely fair comparison. For this reason, each of the three methods is also tested on the out-of-sample test set. These results can be found in 8.

Table 8: Out of-sample-model performance for the benchmark model using three different methods to deal with the unbalanced data. The naive model does not account for the unbalanced data, the weighting method assigns a relatively high weight to the minority class and the Categorical Smote model combines the adjusted version of the SMOTE oversampling method with the weighting method.

	naïve	weighting	Categorical SMOTE
ROC AUC	0.80	0.80	0.80
Brier score	0.00	0.20	0.00
correct positive	0.00%	71.86%	33.90%
false-positive	0.00%	27.50%	13.80%
accuracy	99.96%	72.49%	86.17%
F1	NaN	0.002	0.002

The out-of-sample performance remained roughly the same for the naive approach and the weighting approach suggesting that it is a stable estimation model. These results are not unexpected for the naive approach since it is a model that almost only predicts non-default cases returning the same results independently from the data set used.

The Categorical SMOTE method failed to keep its performance out-of-sample. The correctly identified default cases decreased from 72% to 34%, while the other measures increased in performance. However, the improvement of the other measures does not weigh up against the decrease in the predictive power of the default cases. Based on the out-of-sample performance, the Categorical SMOTE method is able to identify some of the default cases but still prefers to identify most cases as non-default like the naive methodology does.

Whereas the performance of the Categorical SMOTE and the weighting method were comparatively similar using the in-sample data set, the weighting method outperforms the Categorical SMOTE method on the out-of-sample data set. Based on the results it can be concluded that the weighting method is the optimal method to deal with the unbalanced data set. This method will be used as balancing method for the remaining paper.

5.3.3 Model performance in 2020

To further prove the existence of the problem, the Logistic Regression model will be used. This will be done by splitting the data into three groups as shown in figure 7.

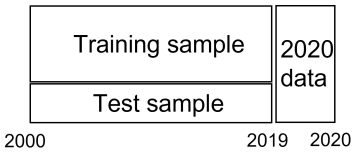

Figure 7: Graphical explanation of how the data will be split to test whether the model performance decreased in 2020. The figure shows how the data for the year 2020 will be set apart and splits the remaining data from 2000 until 2019 into a training set and a test set containing 80% and 20% of the data respectively.

The first data set will contain all observations for the year 2020. The remaining data will be split into a training set containing roughly 80% of the data from 2000 until 2019, and a test set containing the remaining 20% of the data from 2000 until 2019. This way, the model will be estimated and tested on data sets that do not include any data that is collected during the COVID-19 pandemic and will then be tested whether the model still holds its performance in case it is used on the data of the COVID-19 pandemic.

Using this method of estimating the benchmark model, the assumption that the data of 2020 is different compared to the periods before can be confirmed. If the data would behave the same over all the years, the test set with the data of 2020 would give similar results compared with the out-of-sample test set. However, based on the assumption of a change in the behaviour of the data as a result of the COVID-19 pandemic, the model is expected to give worse results for the year 2020 since some of the financial changes in 2020 are not expected to be captured by the data used in the model.

Table 9 contains the confusion matrix for the training set, test data set and the data set with the data for the year 2020. A confusion matrix can be used in a classification model to show the accuracy of the prediction as explained in section 4.7. Table 10 contains some performance statistics corresponding to each of the data sets.

Table 9: Confusion matrices for the benchmark model using the training data set, test data set and data set containing the data for 2020.

	0	1
0	4202438	1314188
1	658	1707

	0	1
0	1049717	329420
1	167	444

	0	1
0	2604	6580
0	9	56

Table 10: Performance measures for the benchmark model for the training data set, test set and the data for year 2020.

	training set	test set	2020
roc auc	0.79	0.80	0.69
Brier score	0.18	0.19	0.44
correct positive	72.18%	72.67%	86.15%
false-positive	23.81%	23.88%	71.14%
accuracy	76.18%	76.11%	28.76%
F1	0.002	0.002	0.017

As can be seen in both the confusion matrix and the table with the performance measures, the model behaves roughly the same in both the training set and the test data set. The model manages to get an accuracy of 76% and is able to identify the defaults in 72% of the cases.

The interesting part is when looking at the out-of-sample test with the data for the year 2020. Applying the estimated model on the data for 2020 returns a correct default prediction of 86% and an increase in the F1 value from 0.002 to 0.017. However, the model seems to identify a lot of the observations as default and identifies 71% of the total default predictions incorrectly. This results in a decrease in accuracy from 72% to only 29%. The model is able to correctly identify a large number of defaults but this is at the cost of an increase in incorrect default predictions and a relatively small number of correct non-defaults.

Based on these results, it can be concluded that the model is no longer able to properly function with the data for the year 2020. Using the financial data collected in 2020, the model expects a larger number of defaults compared to the actual number of defaults, Confirming the assumption that the economy changed in 2020 in such a way that the model is no longer able to function properly.

5.3.4 Complete benchmark model

The complete model will be estimated using all the data available. This data will be split up into a training part using 80% of the data and a test set with the remaining 20%. The results of the full model will be used as benchmark results for the other models to compare.

The performance measures of the benchmark model can be found in table 11. The corresponding confusion matrix is located in table 12.

The model seems overall to perform rather consistent. Both the performance measures based on the training data set and the test data set result in roughly the same values. The biggest difference can be found in the percentage of correctly identified defaults with a decrease from

Table 11: Performance measures for the benchmark model using 80% of the data as training set and the remaining 20% as test data set.

	training set	test set
roc auc	0.81	0.80
Brier score	0.21	0.20
correct positive	73.81%	71.86%
false-positive	27.50%	27.50%
accuracy	72.50%	72.49%
F1	0.002	0.002

Table 12: Confusion matrices for the training data set and the test data set for the benchmark model.

	0	1
0	4004565	1519374
1	642	1809

	0	1
0	1001028	379980
1	166	424

74% to 72% between the training set and the test set. Based on the almost same results on the training and test set, it can be concluded that the benchmark model is a rather consistent model. This consistency is not unexpected since the COVID-19 pandemic data is still only a relative small part of the data and thus the effect is not yet fully visible in the overall model performance.

The estimated coefficients with the biggest impact on the model can be found in table 13. The list of all coefficients can be found in appendix G.

Table 13: The 5 variables with the highest estimated coefficient in the benchmark model. The 'Mining and quarrying' variable is a dummy variable indicating whether the company is active in that industry.

Variable	Coefficient
Tangible net worth	5.28
Total shareholders funds liabilities	-5.21
Solvency	-3.54
Mining and quarrying	2.28
Oher shareholders funds	-2.03

Since each variable has been standardised, the estimated coefficients can be compared. Based on the estimated coefficients, it can be concluded that the Tangible net worth and the Total shareholders funds liabilities variable have the biggest impact on the model, followed by the Solvency. Interestingly is the positive coefficient for the Tangible net worth which suggest that the higher the Tangible net worth of a company is, the bigger the chance that the model will identify that company as default. Tangible net worth can be seen as the total value of the physical assets of a company, and thus a high Tangible net worth would be expected to have a positive effect on the default chance. However, this effect could be explained by looking at the correlation matrix of the variables. Tangible net worth has a correlation value of 0.76 and 0.65

with the variable 'OtherShareholdersFunds' and 'TotalShareHoldersFunds' respectively, as can be seen in appendix B. With this relatively high correlation, the model could use the variables to offset the effect against each other leading to unexpected results.

Lastly, the 'Mining and Quarrying' dummy variable is among the five variables with the biggest estimated coefficients in the model. This is the industry dummy variable indicating that a company active in the Mining and quarrying industry is more likely to default compared to any of the other industries and that this variable has a relatively big impact on the model compared to most of the other variables.

5.4 PSTR model

The PSTR is the first model that is created that should be able to improve the benchmark model. The model is an extension of the Logistic Regression model by adding the variables again but adjusted by an economic indicator as explained in section 4.4.

In section 5.4.1 three different economic indicators will be tested. Each of the economic indicators reacts slightly different to the events that happen in the economy, but these differences could be big enough to give different results when used in the PSTR model. The models with the economic indicators will again be estimated with the data of 2020 as out-of-sample test in a similar fashion as has been done with the benchmark model in section 5.3.3. The economic indicator that results in the best performance measures on the out-of-sample test set will be used to estimate the PSTR model and be compared to the benchmark model.

The full model will be estimated and described in section 5.4.2 using the best performing economic indicator. By comparing it with the benchmark model, a conclusion can be formulated whether the PSTR model is able to outperform the benchmark model, and if so, how much better it is.

5.4.1 Economic indicators

The PSTR model makes use of an economic indicator to adjust the variables accordingly. The economic indicators that will be tested are the GDP, the 'willingness to buy', and the 'consumer confidence' in The Netherlands. The estimation of the model for each of the economic indicators will again be done by splitting the data into three sets. The data are split into a set with the data of 2020 as out-of-sample test and the remaining data is split into 80% and 20% for training and testing purposes respectively.

As was concluded earlier, the benchmark model failed to keep its performance when applied to the data of 2020. If the PSTR model works as expected, the performance of each of the three

models, when applied to the out-of-sample data set of 2020, should perform better compared to the benchmark model. There will likely be differences between the models since different economic indicators are used, but they should not differ too much.

The results of the performances of the models applied to the out-of-sample data set and the data from 2000 until 2019 for each of the three economic indicators are in table 14.

Table 14: performance measures for the PSTR model using GDP, consumer confidence, and willingness to buy as economic indicator.

	GDP		Consumer confidence		willingness to buy	
	test set	2020	test set	2020	test set	2020
Correct positive	79.69%	69.23%	77.46%	73.14%	70.78%	69.07%
False-positive	20.48%	24.43%	22.53%	26.63%	22.49%	22.53%
accuracy	79.51%	65.22%	77.46%	73.14%	77.50%	77.4%
F1	0.003	0.012	0.003	0.034	0.003	0.034

The performance of each of the three models decreased when applied to the data of 2020 compared to the test data set. However, the decrease is not as bad as in the benchmark model. Whereas the accuracy dropped by nearly 50% in the benchmark model, the accuracy decreased by 14% in the PSTR model using the GDP as economic indicator and remained the same using the willingness to buy variable. GDP as economic indicator seems to give the best results based on the test data set but loses its power when applied to the data of 2020. Similarly, the results of the model using the Consumer Confidence variable decreases in performance when comparing the test data set and the data of 2020. Only the Willingness to buy variable manages to keep its performance on the data set of 2020, making it a good and robust model. The results are a bit lower on the test data set but the consistent performance on the data of 2020 makes it the most attractive variable and will thus be used as economic indicator in the estimation of the PSTR model.

5.4.2 Complete PSTR model

To be able to conclude whether the PSTR model is able to outperform the benchmark model, the complete PSTR model needs to be estimated on the whole data set. The model is estimated on 80% of the data and the remaining 20% is used to verify the performance of the model. The economic indicator that is used is the 'Willingness to buy' as explained in section 5.4.1. The results can be found in the confusion matrix in table 16 and the corresponding performance measures are in table 15.

The performance measures of the training data set return comparatively results to the test

Table 15: Performance measures for the PSTR model estimated using 80% of the data. The remaining 20% is used as out-of-sample test set.

	training set	test set
roc auc	0.77	0.75
Brier score	0.19	0.19
correct positive	69.56%	67.97%
false-positive	20.72%	20.69%
accuracy	79.27%	79.29%
F1	0.003	0.003

Table 16: Confusion matrices for the training data set and test data set for the PSTR model.

	0	1
0	4379169	1144770
1	746	1705

	0	1
0	1095123	285885
1	189	401

set suggesting that it is a rather consistent model. Compared to the benchmark model, the PSTR model is slightly worse in correctly identifying defaults with a correct prediction of 68% compared to 72% on the test data set. However, the benchmark model identifies 28% of the data incorrectly as default while the PSTR model only does this to 21%. These results are therefore partly the reason why the overall accuracy and the F1 score of the PSTR model are better than the benchmark model. Whereas the benchmark model had an accuracy of 72% the PSTR model got an accuracy of 79%. The slightly worse default identification power is therefore compensated by relatively better results on the other performance measures.

Overall, the PSTR model improves the results both in the complete model and when applied to the data of 2020. It can thus be concluded that the PSTR model is indeed able to outperform the benchmark model and is better able to deal with the data collected during the COVID-19 pandemic.

The five variables with the highest estimated coefficient are given in table 17. A full list of the coefficients can be found in appendix H.

Table 17: The 5 variables with the highest estimated coefficient in the PSTR model. The variables with 'Adjusted' in their name are the variables that are adjusted by the economic indicator variable.

Variable	Coefficient
Total Shareholders Funds Liabilities Adjusted	-7.62
Other Current Assets Adjusted	-5.69
Cash-Cash equivalent adjusted	-3.84
Tangible Net Worth	3.39
Other Current Assets	-3.28

Interestingly is that the three variables with the highest coefficient are the variables which are adjusted by the economic indicator variable (indicated by the 'adjusted' term in the variable

name). This means that the model assigns the biggest impact on the model to the variables which are added by the PSTR model instead of the original variables. If the model would assign only a low value to the adjusted variables, the performance measures of the PSTR would return more comparable results to the benchmark model. Furthermore, the Tangible net worth variable is again among the variables with the highest coefficient suggesting that this variable could be one of the most important variables from the original data set.

It should be noted that the coefficients in the PSTR model are not as easy to compare to each other as in the benchmark model. Both the benchmark model and the PSTR model used standardised variables. However, the PSTR model scales some of those standardised variables by use of the transition function. Since this transition function is bounded between 0 and 1, it essentially only adds a percentage of the variable depending on the state of the economy. This effect can be seen in the values of the estimated coefficients in which the adjusted variables receive a relatively larger coefficient. The true importance of each variable can therefore no longer be based merely on the value of the coefficient. Nevertheless, the adjusted variables remain important in the PSTR model.

The PSTR model manages to keep its performance when applied to the data collected during the COVID-19 pandemic which the benchmark model failed to do. The PSTR model also outperformed the benchmark model using the whole data set, but the improvement was not as big. However, since the amount of data collected during the COVID-19 pandemic is currently only a relatively small part of the whole data set, the performance could even further increase over time. If the benchmark model does fail to deal with the COVID-19 pandemic data, as the results suggest, the performance of the benchmark model will likely decrease even further as the amount COVID-19 pandemic data increases. This would further strengthen the conclusion that the PSTR model outperforms the benchmark model and would strengthen the advice to use the PSTR model even more.

5.5 Hybrid model

The hybrid model is the last model that will be created. It requires more steps to come to the final model and will therefore be split up into three subsections. The first step is to create a Random Forest model which will be discussed in section 5.5.1. Then, in section 5.5.2, the results of the SHAP importance values will be discussed which will then be used to find the optimal combination of features that are the most important according to the Random Forest. In section 5.5.4 the final results of the hybrid model will be discussed.

5.5.1 Random Forest

The Random Forest is a machine learning technique that is well known to have good performances. One of the biggest downsides of using the Random Forest is that it results in a black box. Due to the regulatory requirements that the decisions need to be explainable, the Random Forest is not commonly used as a credit risk model. The hybrid model uses the Random Forest as initial step to find the most important variables. This section will shortly describe and discuss the results the Random Forest model gives but will not go into too much detail since it is only the first step in the hybrid model.

The Random Forest will be calibrated on the same data set as the other models. The variables given as input for this model are the same as for the benchmark model. The Random Forest requires some hyperparameters to be set before the model can be estimated. These hyperparameters will be set using a grid search on a set of random possible parameters. Using cross validation, this grid search results in an optimal combination of parameters. A table with all tested hyperparameters can be found in appendix I. The optimal combination of parameters will be based on the combination that offers the highest F1 score.

Using the optimal hyperparameters, the Random Forest model is estimated. The results are in the confusion matrix in table 19 and some performance measures for the training set and the out-of-sample data set are given in table 18.

Table 18: In-sample and out-of-sample performance statistics for the Random Forest model using 80% and 20% of the data respectively.

	training set	test set
ROC AUC	0.91	0.83
brier score	0.11	0.11
Correct Positive	79.48%	63.00%
false-positive	13.85%	13.79%
accuracy	86.15%	86.20%
F1	0.005	0.004

Table 19: Confusion matrix for the training set and test set for the Random Forest model.

	0	1
0	4728700	760316
1	499	1933

	0	1
0	1182991	189264
1	225	383

As expected the performance measures of the Random Forest model are indeed better compared to the other models. The model accuracy increases from 73% to 86% while the false-positives decrease from 27% to 14% compared to the benchmark model. Interestingly, the correct positive prediction does decrease quite a bit in the out-of-sample test from 79% to 63%. But, since only the correct default prediction decrease a bit, while the other measures remain the same, the Random Forests model is still deemed to be a good model to be used for the

hybrid model.

Overall, the results of the Random forest are an improvement compared to the benchmark model. This is in line with the expectation that the Random Forest model should be able to outperform a Logistic Regression model.

5.5.2 SHAP

The second step in the hybrid model is to extract the most important variables out of the Random Forest model. This will be done using the SHAP importance values. The SHAP values are calculated by going through each observation and calculating the impact of each variable is on the outcome of the model. Visually, this results in a graph as can be seen in figure 8.

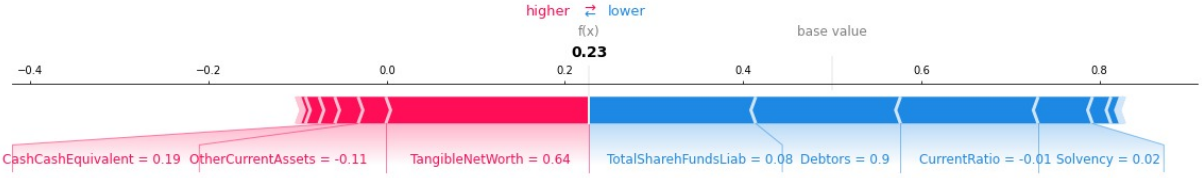

Figure 8: SHAP graph for one observation. The graph shows what decisions the Random Forest made and the impact each variable had on the outcome.

The graph gives an example of how the SHAP values are calculated for one specific observation out of the data set. The model starts each time with a base value, which is comparable to an intercept in a regression model. The base value in this model is around 0.5. The X-axis in the graph indicates the probability that this observation belongs to a default. If the outcome is above 0.5 the observation will be classified as default and non-default otherwise. The model returns a final value of 0.23 for the observation used in this example, and would thus classify it as non-default.

The methodology then continues by calculating the impact of each variable on the outcome. Each of the bars in the graph indicates the impact of one of the variables. The variables that cause an increase in the outcome are in red while the variables that decrease the final outcome are given in blue. The width of each bar shows the magnitude of the effect the corresponding variable had. By doing this, the interpretation and the reasoning why the model gave a specific outcome becomes more clear since the impact of each variable can be explained and described.

By repeating this for each observation, a list with the impact of each variable on each observation can be constructed. By taking the sum of the impact each variable has in each observation, a ranking can be made indicating what variable had the biggest overall impact on

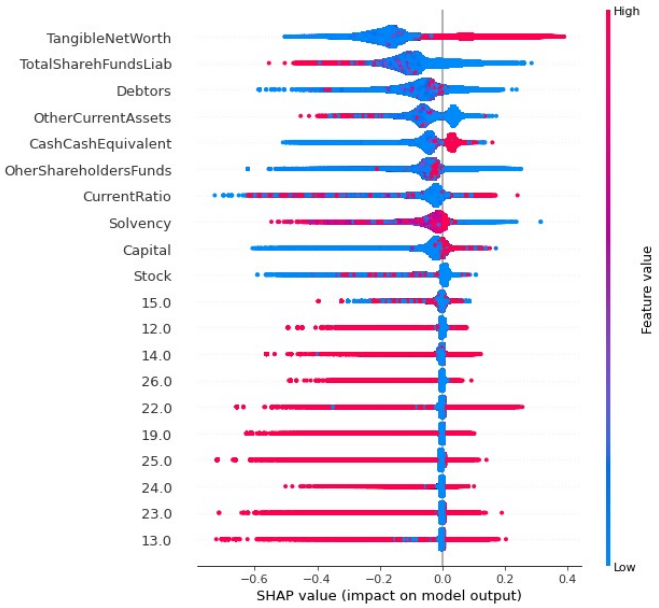

Figure 9: SHAP importance values for 20 variables with the most important variable on top. Each dot in the figure represents an observation in the data set with blue indicating a low observed value and red a high observed value. The X-axis indicate the SHAP value corresponding to each observation.

the model.

Figure 9 is constructed by combining the impact of each observation as just explained. The Y-axis in the graph indicates the 20 variables with the higest SHAP importance values in decreasing order, with the variable with the biggest total impact value on top and the variable with the least impact on the bottom. Each dot represents a single observation, with the corresponding SHAP value on the X-axis. The actual values the variable have are indicated using a colour scale going from blue for relative a low value to a red dot for observations with high values.

By using this graph, the Random Forest model can relatively easily be interpreted. Purely based on the graph, it can be concluded that, for example, the Tangible net worth variable has the biggest impact on the model. Furthermore, a high Tangible net worth increases the chance that the company will be marked as a default, while a low Tangible net worth increases the chance to be classified as a non-default. This result is in line with the results of both the benchmark model and the PSTR model which had the Tangible net worth variable among the most important variables and also showed that a high Tangible net worth increases the result to be marked as default.

The SHAP values for the various industries is pretty interesting [3]. It should be noted that the industry variables are dummy variables being either a value of 1 if the company is active

[3]Due to the method that is used to estimate the models and calculate the SHAP values, each industry is replaced by a number. A list of corresponding industries to each number is located in appendix D.

in that industry and a 0 otherwise. The graph suggests that the Random Forest model assigns a relatively big impact to being active in a specific industry. Looking ,for example, at the 'Manufacturing'(number 25.0 in the graph) industry, some of the observations received a SHAP value slightly below -0.6 while the highest SHAP value is only 0.2, suggesting that the Random Forest assigned quite a positive impact to that variable to be classified as non-default.

The Random Forest model seems to assign quite a large part of the default explanation to the industry the company is active in. This effect could be due to the rather limited number of used explanatory variables resulting in the model assigning the remaining effect to the different industry dummies and thus using it as a sort of error term.

5.5.3 Interaction terms

The SHAP methodology is able to calculate the interaction effect as explained in section 4.6.1. The SHAP interaction value is the feature effect of the combined variables after subtracting the individual feature effect. This method calculates whether there exists a combination of features that have a high impact on the Random Forest model. It could be the case that the Random Forest model found certain combinations of features to be important that were not accounted for in the benchmark model.

The graph in figure 10 contains the 18 interacting variables with the biggest average SHAP value. On the Y-axis are the various interacting variables in increasing order. While the X-axis has the corresponding SHAP value. Based on the graph, the variables with the highest interaction SHAP value can be added in the last step of the hybrid model.

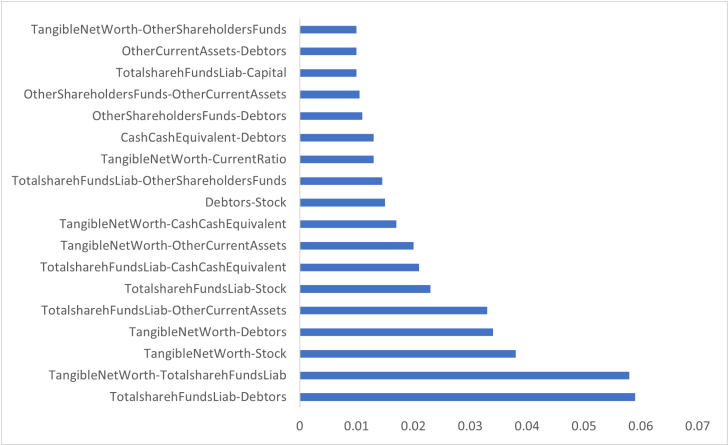

Figure 10: SHAP importance values for 18 interaction variables with the most important variable on the bottom.

5.5.4 Final model

The last step of the hybrid model combines the previous steps by selecting the variables with the highest SHAP importance values and adding them to the PSTR model. If the performances of the hybrid model are able to outperform the PSTR, it can be concluded that some of the important information that was present in the data was not yet captured by the PSTR model.

For the model estimation, the six interaction variables with the highest SHAP value are added to the PSTR model. The results of the model estimation, after adding the six interaction variables, are given in the confusion matrix in table 20 and the corresponding performance measures are given in table 21.

Table 20: Confusion matrices for the training data set, test data set and data set containing the data for 2020 based on the hybrid model.

	0	1
0	4381324	1135302
1	743	1622

set	0	1
0	1094468	284669
1	182	429

	0	1
0	6682	2502
0	19	46

Table 21: Performance measures for the training set, test set, and data set using the data for the year 2020. The performance measures are based on the results of the hybrid model.

	training set	test set	2020
roc auc	0.77	0.75	0.72
Brier score	0.19	0.19	0.22
correct positive	68.58%	70.21%	70.77%
false-positive	20.57%	20.53%	27.05%
accuracy	79.42%	79.35%	72.74%
F1	0.003	0.003	0.035

The performance measure on the hybrid model based on the training set and the test set is similar or slightly worse on almost all of the measures compared to the PSTR model, suggesting that the hybrid model is performing worse. The hybrid model has similar performances on the data of 2020, except for the false-positive classification and the accuracy. The PSTR model had an accuracy of 77% on the data of 2020, the hybrid model only managed to get an accuracy of 73%. Similarly, the false-positive classification in the hybrid model are worse, with 27% false-positive against 23% false-positive in the PSTR model. It can thus be concluded that the hybrid model is not better able to deal with the data of 2020 compared to both the PSTR. Apparently, the interactions between the variables lacks information to be able to further improve the PSTR model.

5.5.5 Complete hybrid model

Finally, the hybrid model is estimated using all the available data. The data set is again split into a data set containing 80% of the data for model training and the remaining data as test set. The results are given in the confusion matrix in table 23 and the corresponding performance measures are given in table 22.

Table 22: In-sample and out-of-sample performance statistics for the hybrid model using 80% and 20% of the data respectively.

	training set	test set
ROC AUC	0.76	0.74
brier score	0.19	0.19
Correct Positive	69.11%	67.97%
false-positive	20.60%	20.58%
accuracy	79.39%	79.41%
F1	0.003	0.003

Table 23: Confusion matrix for the training set and test set for the hybrid model.

	0	1
0	4385454	1138485
1	757	1694

	0	1
0	1096694	284314
1	189	401

Based on the performance measures of both the hybrid model and the PSTR model, it can be concluded that the hybrid model fails to further improve the model performance. Both in the training set as well as in the test set, the differences in the model performance is neglectable. This means that the addition of the interaction terms does not bring any noteworthy additional power to the PSTR model. Since the hybrid model requires more steps to be estimated and uses more variables in the final model, it is not desirable to use the hybrid model compared to the PSTR model.

A table with the five variables with the highest estimated coefficient is given in table 24.

Table 24: The 5 variables with the highest estimated coefficient in the hybrid model. The adjusted variables are the variables that are adjusted by the economic indicator variable.

Variable	Coefficient
Tangible Net Worth	3.22
Other Current Assets Adjusted	1.75
Total Shareholders Funds Liabilities	-1.60
Other Shareholders Funds	-1.55
Total Shareholders Funds Liabilities Adjusted	-1.46

The five variables with the highest coefficient confirm the earlier mentioned conclusion that the addition of the interaction variables is neglectable since none of the interaction variables is among the highest estimated coefficient. The interaction variable with the highest coefficient is

ranked place 20 of the highest coefficients with a value of -0.31. This means that the effect of these new variables is relatively small. Combined with the performance measures on the data of 2020, it can be concluded that the hybrid model fails to further improve the PSTR model and even decreases the model performances in some cases.

6 Conclusion

As a result of the start of the COVID-19 pandemic in early 2020, the economy in The Netherlands, and around the world, has changed. Some industries were completely shut down or were no longer able to function properly. At the same time, banks and governments took action to support those companies and to prevent them from going into default. As a consequence, companies that were set to default according to the credit risk models, might still be standing making the models obsolete.

This research aims to investigate whether the default predictions of the current credit risk models are indeed worse since the start of the COVID-19 pandemic. Then the research sets out to find a new model to ensure that the data collected during the COVID-19 pandemic remains usable. This will especially be important in the coming years in which the banks use historical data that contains the data collected during the COVID-19 pandemic for their credit risk model.

The models will be estimated using a database containing financial information for various companies across 19 different industries. In the paper, it was established that the change in financial data was significantly different going into 2020 compared to the years before. The change in data, between 2019 and 2020, for 5 variables was significantly different compared to the years before. This suggests that there indeed has been a change in the behaviour of the financial data of the companies in 2020.

In total, four different models were estimated in this paper. A Logistic Regression model was created as a benchmark model, after which a Panel Smooth Transition Regression model and a Random Forest model were estimated as potential improvements over the benchmark model. Since the Random Forest results in a rather difficult to understand model, this model was extended into a hybrid model to improve interpretability by the use of SHAP importance values.

Each of the models was estimated twice. In the first estimation, the data was split up into three parts in which the first set contained all the data of 2020. The remaining data was split up into a training data set and a test data set using 80% and 20% of the data respectively. This method uses the data of 2020 as an out-of-sample test and can be used to verify the performance of the models in 2020 while not being calibrated using any of the data collected in 2020.

To establish a baseline to evaluate whether the new models are able to improve the problem, a Logistic Regression model was created as a benchmark. A Logistic Regression model is a relatively easy model and is widely used in the banking industry as credit risk models, making it a good benchmark model. The performance of the benchmark model decreased quite rapidly when applied to the data of 2020 compared to the years before. The benchmark model was able to correctly identify 73% of the defaults with a model accuracy of 76%. However, when the data for 2020 was used in this model, the percentage of correctly identified defaults increased to 86% but at the cost that the accuracy fell to 29%. This was partly due to the model identifying 72% of all the default classification incorrectly.

The second model was the Panel Smooth Transition Regression model. This model essentially copies all the existing variables in the Logistic Regression models and adds them again into the model but adjusted to the state of the economy in that year. Using the initial model estimate method with the data of 2020 as test data, the performance did not change that much. The correctly identified defaults fell from 71% in the test data set to 69% in the data of 2020, while the incorrect default indentifications remained at around 22%. This suggests that the Panel Smooth Transition Regression model is able to deal with the COVID-19 pandemic data and that the model performance remained stable.

The Random Forest model was estimated using the same set of variables as the benchmark model. The performance of the Random Forest model was a big increase compared to the benchmark model with correct default prediction of nearly 80% in-sample and an incorrect default prediction percentage of 14%. However, the correct default prediction did decrease quite a bit, from 80% to 63% when applied to the test data set. Since the performances remained stable between the training data set and the test data set, except for the correct positive predictions, the Random Forest model seemed to be an improvement over the benchmark model. The improvement of the Random Forest model is not unexpected since the Random Forest is known to reach high accuracy and good model performance. However, the downside is the limited interpretability of the Random Forest model which is also the reason why the model is not yet widely used.

The final model, the hybrid model, combines the power of the Random Forest model with the interpretability of the PSTR model. This is achieved by using SHAP importance values. The SHAP importance values are calculated based on the relative impact each variable had in the Random Forest model. Besides the impact of each variable seperately, the impact of each combination of two variables is also calculated. This results in an extensive list ranking each combination and individual variable from most important to least in the Random Forest model.

The combination of variables with the highest importance value was added to the PSTR model to possibly improve the PSTR model. The hybrid model managed to return similar results to the PSTR model based on the whole data set. However, when the data was split into seperate data sets with the data of 2020 as a seperate test set, the performance of the hybrid model decreased. The Accuracy of the hybrid model decreased from 79% to 73% once it was applied to the data of 2020, while the incorrect default predictions increased from 21% to 27%. This means that the hybrid model failed to bring any improvement to the PSTR model and that the hybrid model was even worse at dealing with the data collected in 2020 during the COVID-19 pandemic.

Both the PSTR model and the hybrid model managed to outperform the Logistic Regression model. However, only the PSTR model managed to keep its performance once it was applied to the COVID-19 data. The hybrid model was stil able to outperform the benchmark model but did not perform as well as the PSTR model did. It can be concluded that the existing credit risk models can be improved to deal with the COVID-19 pandemic data using the proposed methodologies. However, based on the complexity of the hybrid model and the similar or worse performances compared to the PSTR model, the PSTR model is the recommended model with the advantage that it is only a relatively easy modification to existing models and that the credit risk model can thus easily be adapted to deal with the COVID-19 pandemic.

7 Discussion

This paper has investigated whether the impact of the COVID-19 pandemic on the credit risk models could be incorporated in the models. Both the PSTR model and the hybrid model were able to be an improvement compared to the Logistic Regression model. The PSTR model was even able to outperform the hybrid model based on some of the performance measures. Since the COVID-19 pandemic is not yet over at the time of writing, it is not yet possible to reach a complete conclusion. However, the results suggest that model improvements are possible and that the improvements may be able to properly deal with the data of the COVID-19 pandemic.

The main contribution of this paper is to primarily investigate whether the impact of COVID-19 pandemic is already visible in the financial credit risk data. Furthermore, the paper creates a starting point for further research for possible solutions to properly deal with the data of the COVID-19 pandemic in credit risk models. Based on the currently available data, it can already be concluded that the financial data seem to behave significantly different compared to the years before. Both the PSTR model and the hybrid model were relatively better able to capture the changes in data compared to the Logistic Regression model and they are therefore more useful

to use in the years to come after the pandemic.

In the coming periods, more data will become available resulting in more accurate results. Further research can continue on the methodology set out in this paper using the additional data to get to a stronger conclusion of the power of the PSTR model and the hybrid model.

Using the results of this paper, that improvements are possible to deal with the COVID-19 pandemic data, it can be of interest to focus on the more advanced credit risk models that some banks use. This paper used a Logistic Regression model as a benchmark to ensure that the results are relevant for all types of banks, whether they use simple models or more complicated models. The simplicity of the PSTR modelling technique ensures that the method can be applied to a wide range of models since it essentially copies the existing model but is adjusted to the state of the economy. Continuing on this research by adjusting the more advanced models could therefore lead to even better results.

Due to the limited amount of data available, the model was calibrated using 7 variables and 3 financial ratios. This resulted in a benchmark model whose performances were not optimal. However, since the suggested models still were able to bring improvements to the benchmark model using the limited data, using more relevant financial ratios and variables could likely lead to an even better improved model.

The COVID-19 pandemic is quite a unique crisis since the financial impact is not equal in every industry. Industries such as tourism were shut down almost completely while online retail might even have better results than the years before. Investigating whether this difference does exist and how big the difference is could be an interesting next step. If the impact of the industries is indeed differently, creating credit risk models depending on their respective industries could improve the performances even more.

References

Aas, K., Jullum, M. and Løland, A. (2019), 'Explaining individual predictions when features are dependent: More accurate approximations to shapley values', *arXiv preprint arXiv:1903.10464* .

Allen, M. (1997), 'Understanding regression analysis: The problem of multicollinearity in:understanding regression analysis', pp. 176–180.
URL: *https://doi.org/10.1007/978-0-585-25657-3*$_3$7

Ballentine, C., Ponczek, S. and Hajric, v. (2020), 'S&p 500 plunges 7%, triggering market-wide stock trading halt', *Bloonberg* .
URL: *https://www.bloomberg.com/news/articles/2020-03-08/rout-in-u-s-stock-futures-would-trigger-trading-curbs-at-5*

Baltagi, B. et al. (2008), *Econometric analysis of panel data*, Vol. 4, Springer.

Beaver, W. (1966), 'Financial ratios as predictors of failure', *Journal of Accounting Research* **4**, 71–111.

Belgiu, M. and Drăguţ, L. (2016), 'Random forest in remote sensing: A review of applications and future directions', *ISPRS journal of photogrammetry and remote sensing* **114**, 24–31.

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**, 5–32.

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002), 'Smote: synthetic minority over-sampling technique', *Journal of artificial intelligence research* **16**, 321–357.

Couronné, R., Probst, P. and Boulesteix, A.-L. (2018), 'Random forest versus logistic regression: a large-scale benchmark experiment', *BMC bioinformatics* **19**(1), 1–14.

Daoud, J. (2017), 'Multicollinearity and regression analysis', *Journal of Physics: Conference Series* **949**.

Dong, G., Lai, k. K. and Yen, J. (2012), 'Credit scorecard based on logistic regression with random coefficients', *Procedia Computer Science* **1**, 2463–2468.

Fryer, D., Strümke, I. and Nguyen, H. (2021), 'Shapley values for feature selection: the good, the bad, and the axioms', *arXiv preprint arXiv:2102.10936* .

Ghosh, D. and Vogt, A. (2012), 'Outlier: An evaluation of methodologies', *Joint Statistical meetings* .

González, A., Teräsvirta, T., van Dijk, D. and Yang, Y. (2017), Panel smooth transition regression models, SSE/EFI Working Paper Series in Economics and Finance 604, Stockholm School of Economics.
URL: *https://EconPapers.repec.org/RePEc:hhs:hastef:0604*

Goodwin, B., Holt, M. and Prestemon, J. (2011), 'north american oriented strand board markets, arbitrage activity, and market price dynamics: A smooth transition approach', *American Journal of Agricultural Economics* **93**, 993–1014.

Hoerl, A. E. and Kennard, R. W. (1970), 'Ridge regression: Biased estimation for nonorthogonal problems', *Technometrics* **12**(1), 55–67.

Hurn, A., Silvennoinen, A. and Teräsvirta, T. (2014), 'A smooth transition logit model of the effects of deregulation in the electricity market', *NCER Working Paper Series* **Working Paper 100**.

Jakobsen, J., Gluud, C., Wetterslev, J. and Winkel, P. (2017), 'When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide with flowcharts', *BMC Med Res Methodol* **17**, 162.

Jeni, L. A., Cohn, J. F. and De La Torre, F. (2013), 'Facing imbalanced data–recommendations for the use of performance metrics', *2013 Humaine association conference on affective computing and intelligent interaction* pp. 245–251.

Johnson, D. R. (1995), 'Alternative methods for the quantitative analysis of panel data in family research: Pooled time-series models', *Journal of Marriage and the Family* **57**, 1065–1077.

Kadilli, A. and Markov, N. (2012), 'A panel smooth transition regression model for the determinants of inflation expectation and credibility in the ecb and the recent financial crisis', *Available at SSRN: https://ssrn.com/abstract=1903853* .

Khandani, A. E., Kim, A. J. and Lo, A. W. (2010), 'Consumer credit risk model via machine-learning algorithms', *Journal of banking & Finance* **34**, 2767–2787.

Kim, T. (2015), 'T test as a parametric statistic', *Korean Journal of Aneshtesiology* **68**(6), 540–546.

Koenker, R. and Bassett, G. J. (1978), 'Regression quantiles', *Econometrica* **46**(1), 33–50.

Kruppa, J., Schwarz, A., Arminger, G. and Ziegler, A. (2013), 'Consumer credit risk: Individual probability estimates using machine learning', *Expert Systems with Applications* **40**, 5125–5131.

Kuan, C.-M., Michalopoulos, C. and Xiao, Z. (2013), 'Quantile regression on multiple quantile ranges', **Available at SSRN: https://ssrn.com/abstract=1869379 or http://dx.doi.org/10.2139/ssrn.1869379**.

Lessmann, S., Baesens, B., Seow, H.-v. and Thomas, L. C. (2015), 'benchmarking state-of-the-art classification algorithms for credit scoring: An update of research', *European Journal of Operational Research* **247**, 124–136.

levy, J. J. and James O'Malley, A. (2020), 'Don't dismiss logistic regression: the case for sensible extraction of interactions in the era of machine learning', *BMC Medical Research Methodology* **20**, 171–186.

Lundberg, S. M. and Lee, S.-I. (2017), A unified approach to interpreting model predictions, *in* 'Advances in Neural Information Processing Systems', Vol. 30, Curran Associates, Inc. **URL:** *https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf*

Mansfield, E. and Helms, B. (1982), 'Detecting multicolinearity', *The American Statistician* **36**, 158–160.

Mirza, N., Rahat, B., Naqvi, B. and Rizvi, S. K. A. (2020), 'Impact of covid-19 on corporate solvency and possible policy responses in the eu', *The Quarterly Review of Economics and Finance* **available: https://doi.org/10.1016/j.qref.2020.09.002**.

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A. and Brown, S. D. (2004), 'An introduction to decision tree modeling', *Journal of Chemometrics: A Journal of the Chemometrics Society* **18**(6), 275–285.

Osborne, J. (2004), 'The power of outlier (and why researchers should always check for them', *Practical Assesment, Research, and Evaluation* **9**(1), 6.

Pal, M. (2007), 'Random forest classifier for remote sensing classification', *International Journal of Remote Sensing* **26**, 217–222.

Pereira, S., Meier, R., McKinley, R., Wiest, R., Alves, V., Silva, C. A. and Reyes, M. (2018), 'Enhancing interpretability of automatically extracted machine learning features: application to a rbm-random forest system on brain lesion segmentation', *Medical image analysis* **44**, 228–244.

Rad, E., Vahedi, S., Teimourizad, A., Esmaeilzadeh, F., Hadian, M. and Pour, A. (2013), 'Comparison of the effects of public and private health expenditures on the health status: a

panel data analysis in easter mediterranean countries', *International journal of health policy and management* **1**(2), 163.

Randewich, N. (2020), 'Coronavirus, oil collapse erase $5 trillion from u.s. stocks', *Reuters* .
URL: *https://www.reuters.com/article/us-health-coronavirus-stocks-carnage/coronavirus-then-oil-collapse-erase-5-trillion-from-u-s-stocks-idUSKBN20W2TJn*

Schmidheiny, K. (2021), 'Panel data: Fixed and random effects', *Unversität Basel[Handout]* .

Schroeder, M., Lander, J. and Levine-Silverman, S. (1990), 'Diagnosing and dealing with multicollinearity', *Western Journal of Nursing Research* **12**, 175–187.

Torres-Reyna, O. (2007), 'Panel data analysis fixed and random effects using stata', *Data Statistical Services, Princeton University [Handout]* .

Zhang, Z. (2016), 'Missing data imputation: focusing on single imputation', *Ann Transl Med.* **9**, 4.

Appendix

A Variables

Table 25: Table with all variables present in the initial data set. The variables names are written exactly as they are given in the original data set.

IdNumber	OtherCurrentLiabilities	PLForPeriodNetIncome
FixedAssets	TotalSharehFundsLiab	ExportRevenue
IntangibleFixedAssets	WorkingCapital	MaterialCosts
TangibleFixedAssets	NetCurrentAssets	CostsOfEmployees
OtherFixedAssets	EnterpriseValue	DepreciationAmortization
CurrentAssets	NumberOfEmployees	InterestPaid
Stock	OperatingRevenue	ResearchDevelopmentExpenses
Debtors	Sales	CashFlow
OtherCurrentAssets	CostsOfGoodsSold	AddedValue
CashCashEquivalent	GrossProfit	Ebitda
TotalAssets	OtherOperatingExpenses	CostsOfEmployeesOperatingRevenue
ShareholdersFunds	OperatingPLEbit	AverageCostOfEmployee
Capital	FinancialRevenue	ShareholdersFundsPerEmployee
OtherShareholdersFunds	FinancialExpenses	WorkingCapitalPerEmployee
NoncurrentLiabilities	FinancialPL	TotalAssetsPerEmployee
LongTermDebt	PLBeforeTax	SizeClass
OtherNoncurrentLiabilities	Taxation	SecondaryCode
Provisions	PLAfterTax	IndustryLabel
CurrentLiabilities	ExtrAndOtherRevenue	StatusYear
Loans	ExtrAndOtherExpenses	StatusInDefault
Creditors	ExtrAndOtherPL	

B Variable correlation

Table 26: Matrix with the correlation coefficient between each variable that has been used in the model estimation.

	Stock	Debtors	OtherCurrentAssets	CashCashEquivalent	ShareholdersFunds	Capital	OherShareholdersFunds	TotalSharehFundsLiab	CurrentRatio	TangibleNetWorth	Solvency
Stock	1.00	0.36	0.27	0.28	0.16	0.04	0.06	0.29	0.00	0.09	0.00
Debtors	0.36	1.00	0.32	0.33	0.30	0.11	0.22	0.49	0.01	0.24	0.00
OtherCurrentAssets	0.27	0.32	1.00	0.81	0.25	0.08	0.16	0.40	0.00	0.18	0.00
CashCashEquivalent	0.28	0.33	0.81	1.00	0.24	0.09	0.16	0.40	0.00	0.18	0.00
ShareholdersFunds	0.16	0.30	0.25	0.24	1.00	0.24	0.79	0.74	0.01	0.94	0.00
Capital	0.04	0.11	0.08	0.09	0.24	1.00	0.11	0.21	0.00	0.23	0.00
OherShareholdersFunds	0.06	0.22	0.16	0.16	0.79	0.11	1.00	0.54	0.01	0.76	0.00
TotalSharehFundsLiab	0.29	0.49	0.40	0.40	0.74	0.21	0.54	1.00	0.00	0.65	0.00
CurrentRatio	0.00	0.01	0.00	0.00	0.01	0.00	0.01	0.00	1.00	0.01	0.00
TangibleNetWorth	0.09	0.24	0.18	0.18	0.94	0.23	0.76	0.65	0.01	1.00	0.00
Solvency	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

C Data descriptive

Table 27: Some of the data descriptive about the variables that are used in the model estimation before outliers were removed.

	Mean	Standard Deviation	Minimum	Maximum
Stock	615155.13	55109437.42	-5.30E+08	32225000000
Debtors	2360704.11	153165507.11	-2.42E+09	2.02487E+11
OtherCurrentAssets	978642.54	78546203.59	-9.75E+09	1.19783E+11
CashCashEquivalent	1014602.91	47871286.58	-3.93E+08	45462000000
ShareholdersFunds	7313225.17	393705292.55	-5.54E+11	1.51607E+11
Capital	1011495.59	108767440.71	-1.92E+09	1.18697E+11
OherShareholdersFunds	5337186.47	331218083.38	-5.54E+11	1.51109E+11
TotalSharehFundsLiab	14670875.81	563925059.31	-1.33E+10	2.56962E+11
CurrentRatio	371.93	312277.32	-1.94E+08	735499608
TangibleNetWorth	8672.92	402855.39	-5.54E+08	319482224.7
Solvency	-0.519883271	104.6328106	-104948.599	4426.413066

Table 28: Some of the data descriptive of the used variables after the outliers were removed based on a Z-score test.

	Mean	Standard Deviation	Minimum	Maximum
Stock	209037.81	2526506.00	-78237155.00	165919551.00
Debtors	853565.86	7676260.13	-330908000.00	461605193.00
OtherCurrentAssets	317912.56	2813300.99	-115737366.00	236560918.00
CashCashEquivalent	537051.72	2494509.79	-96532000.00	144573712.00
ShareholdersFunds	2059401.29	23945879.22	-1163029689.00	1188398000.00
Capital	388550.99	3899441.76	-294520072.00	327300000.00
OherShareholdersFunds	2355503.41	18443887.15	-980222367.00	998864000.00
TotalSharehFundsLiab	4539712.01	39246804.05	-308807000.00	1706341805.00
CurrentRatio	123.92	5859.73	-930184.00	936746.00
TangibleNetWorth	4270.07	24107.23	-1163029.69	1188398.00
Solvency	-0.04	2.48	-313.44	289.66

D Industries

Table 29: A list of each industry that is present in the data set. Each industry is linked with a unique number that is used in the data set to indicate the specific industry.

11	Construction
12	Professional, scientific and technical activities
13	Accommodation and food service activities
14	Wholesale and retail trade; repair of motor vehicles and motorcycles
15	Financial and insurance activities
16	Transportation and storage
17	Education
18	Arts, entertainment and recreation
19	Real estate activities
20	Agriculture, forestry and fishing
21	Other service activities
22	Administrative and support service activities
23	Human health and social work activities
24	Information and communication
25	Manufacturing
26	Financial and insurance activities
27	Mining and quarrying
28	Water supply; sewerage, waste management and remediation activities
29	Electricity, gas, steam and air conditioning supply

E PSTR grid search

Table 30: All tested combinations for the γ and c parameters in the PSTR model with some of the corresponding performance statistics.

γ	c	correct prediction	false positive	correct positive	precision	recall	accuracy	F1
0.1	-5	0.52335	0.47649	0.61728	0.00055	0.61728	0.52335	0.00110
0.1	-1	0.52418	0.47567	0.62963	0.00056	0.62963	0.52418	0.00112
0.1	0.1	0.52531	0.47453	0.62346	0.00056	0.62346	0.52531	0.00111
0.1	1	0.53159	0.46825	0.62346	0.00056	0.62346	0.53159	0.00113
0.1	5	0.52466	0.47518	0.62346	0.00056	0.62346	0.52466	0.00111
0.5	-5	0.50839	0.49146	0.64198	0.00055	0.64198	0.50839	0.00111
0.5	-1	0.51442	0.48542	0.62346	0.00054	0.62346	0.51442	0.00109
0.5	0.1	0.51654	0.48330	0.62963	0.00055	0.62963	0.51654	0.00110
0.5	1	0.51687	0.48297	0.62963	0.00055	0.62963	0.51687	0.00110
0.5	5	0.50839	0.49145	0.61728	0.00053	0.61728	0.50839	0.00106
1	-5	0.50430	0.49555	0.64815	0.00055	0.64815	0.50430	0.00111
1	-1	0.51858	0.48126	0.62963	0.00055	0.62963	0.51858	0.00111
1	0.1	0.52367	0.47617	0.62963	0.00056	0.62963	0.52367	0.00112
1	1	0.52272	0.47712	0.62346	0.00055	0.62346	0.52272	0.00111
1	5	0.49923	0.50061	0.62963	0.00053	0.62963	0.49923	0.00106
5	-5	0.48341	0.51644	0.64198	0.00053	0.64198	0.48341	0.00105
5	-1	0.53569	0.46415	0.60494	0.00055	0.60494	0.53569	0.00110
5	0.1	0.53531	0.46453	0.61728	0.00056	0.61728	0.53531	0.00112
5	1	0.52861	0.47122	0.60494	0.00054	0.60494	0.52861	0.00109
5	5	0.51379	0.48605	0.61728	0.00054	0.61728	0.51379	0.00108
10	-5	0.48343	0.51642	0.64198	0.00053	0.64198	0.48343	0.00105
10	-1	0.53871	0.46112	0.59259	0.00054	0.59259	0.53871	0.00109
10	0.1	0.53227	0.46756	0.59259	0.00054	0.59259	0.53227	0.00107
10	1	0.52562	0.47421	0.59877	0.00053	0.59877	0.52562	0.00107
10	5	0.51445	0.48540	0.62963	0.00055	0.62963	0.51445	0.00110

F Delta histograms

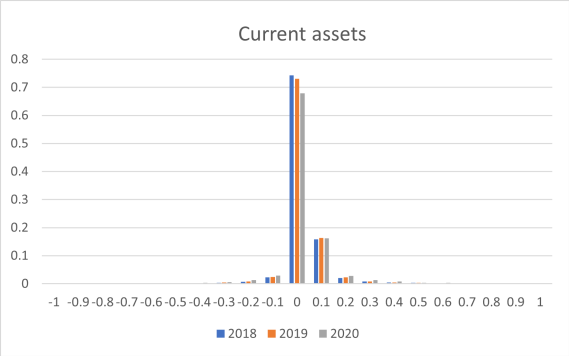

Figure 11: Histogram for the delta values of the Curent assets variable. The y-axis indicates the fractions, being between 0 and 1, of the corresponding variable.

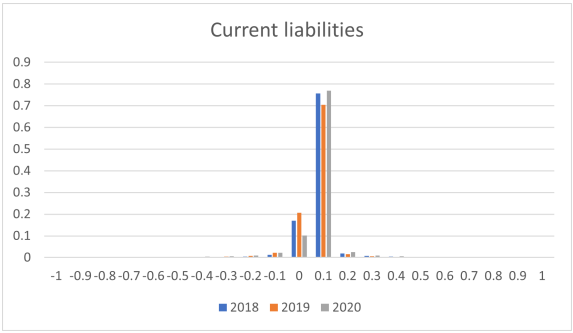

Figure 12: Histogram for the delta values of the Current liabilities variable. The y-axis indicates the fractions, being between 0 and 1, of the corresponding variable.

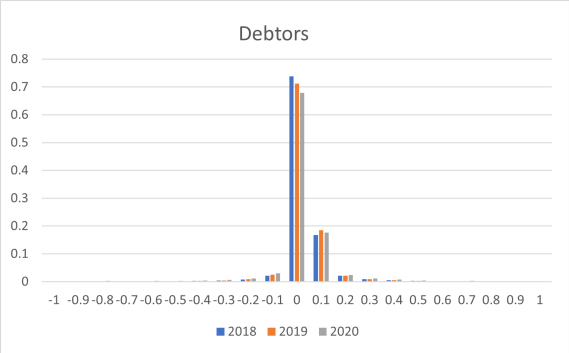

Figure 13: Histogram for the delta values of the Debtors variable. The y-axis indicates the fractions, being between 0 and 1, of the corresponding variable.

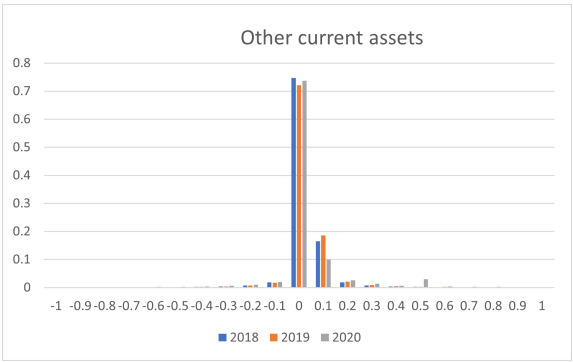

Figure 14: Histogram for the delta values of the Other current assets variable. The y-axis indicates the fractions, being between 0 and 1, of the corresponding variable.

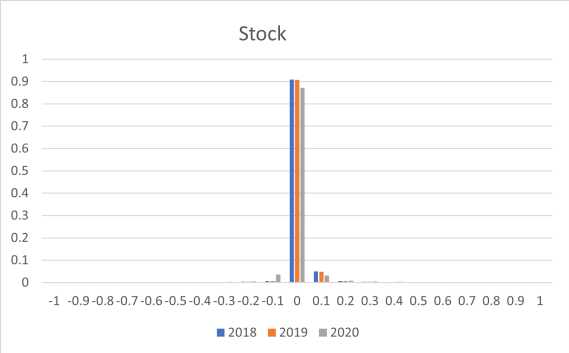

Figure 15: Histogram for the delta values of the Stock variable. The y-axis indicates the fractions, being between 0 and 1, of the corresponding variable.

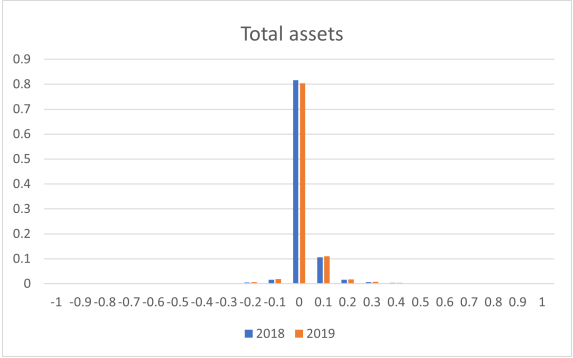

Figure 16: Histogram for the delta values of the Total assets variable. The y-axis indicates the fractions, being between 0 and 1, of the corresponding variable.

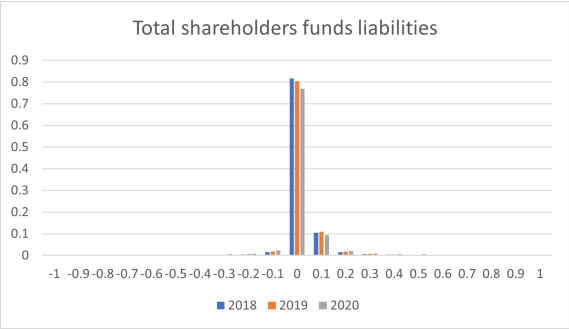

Figure 17: Histogram for the delta values of the Total shareholders funds liabilities variable. The y-axis indicates the fractions, being between 0 and 1, of the corresponding variable.

G Benchmark coefficients

Table 31: The estimated coefficients corresponding to the benchmark model using 80% of the available data. The variables starting with a number are the dummy variables for the specific industries the company is active in.

Variable	Coefficient	Variable	Coefficient
intercept	-4.49	(15) Financial and insurance activities	-0.56
Stock	-0.58	(22) Administrative and support service activities	-0.17
Debtors	0.34	(17) Education	-0.82
OtherCurrentAssets	0.00	(23) Human health and social work activities	-0.55
CashCashEquivalent	-0.04	(20) Agriculture, forestry and fishing	-0.95
Capital	-0.39	(26) Financial and insurance activities	-1.15
OherShareholdersFunds	-2.03	(24) Information and communication	-0.57
TotalSharehFundsLiab	-5.21	(19) Real estate activities	-0.85
CurrentRatio	0.05	(29) Electricity, gas, steam and air conditioning supply	-1.10
TangibleNetWorth	5.28	(18) Arts, entertainment and recreation	-0.33
Solvency	-3.54	(13) Accommodation and food service activities	0.03
(12) Professional, scientific and technical activities	-0.98	(16) Transportation and storage	-0.39
(27) Mining and quarrying	2.28	(21) Other service activities	0.43
(25) Manufacturing	-0.29	(28) Water supply; sewerage, waste management and remediation activities	1.07
(14) Wholesale and retail trade; repair of motor vehicles and motorcycles	-0.30	timevariable	0.31

H PSTR coefficients

Table 32: The estimated coefficients corresponding to the PSTR model using 80% of the available data. The variables starting with a number are the dummy variables for the specific industries the company is active in. The variables that are adjusted by the transition function are indicated by the term 'adjusted' at the end of the variable name.

Variable	Coefficient	Variable	Coefficient
intercept	-0.38	(26) Financial and insurance activities	-0.69
Stock	-0.32	(24) Information and communication	-0.02
Debtors	-0.27	(19) Real estate activities	-0.23
OtherCurrentAssets	-3.28	(29) Electricity, gas, steam and air conditioning supply	-0.47
CashCashEquivalent	2.55	(18) Arts, entertainment and recreation	0.04
Capital	-0.53	(13) Accommodation and food service activities	0.24
OherShareholdersFunds	-2.19	(16) Transportation and storage	-0.39
TotalSharehFundsLiab	-0.41	(21) Other service activities	0.36
CurrentRatio	-0.11	(28) Water supply; sewerage, waste management and remediation activities	0.67
TangibleNetWorth	3.39	StockAdjusted	-0.32
Solvency	0.21	DebtorsAdjusted	1.21
(12) Professional, scientific and technical activities	-0.22	OtherCurrentAssetsAdjusted	5.69
(27) Mining and quarrying	1.35	CashCashEquivalentAdjusted	-3.84
(25) Manufacturing	-0.26	CapitalAdjusted	0.87
(14) Wholesale and retail trade; repair of motor vehicles and motorcycles	-0.20	OherShareholdersFundsAdjusted	2.63
(15) Financial and insurance activities	-0.21	TotalSharehFundsLiabAdjusted	-7.62
(22) Administrative and support service activities	-0.17	CurrentRatioAdjusted	0.72
(17) Education	-0.25	TangibleNetWorthAdjusted	0.34
(23) Human health and social work activities	-0.10	SolvencyAdjusted	-1.20
(20) Agriculture, forestry and fishing	-0.56		

I Random Forest grid search

Table 33: All tested values for the hyperparameters used in the Random Forest. The optimal combination of hyperparameters is based on the combination that results in the highest F1 score.

Parameter	Description	Tested values
Bootstrap	Are the samples bootstrapped	True, False
max depth	The maximum depth of a tree	10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None
max features	The number of features considerd for a split	automatic, squared, log, None
minimum sample leaf	The minimum number of samples required to be at a node	1, 2, 4
minimal sample split	The minimum number of samples required to split an node	2, 5, 10
n estimators	The number of trees used	200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000

J Hybrid coefficients

Table 34: The estimated coefficients corresponding to the hybrid model using 80% of the available data. The variables starting with a number are the dummy variables for the specific industries the company is be active in. The variables that are adjusted by the transition function are indicated by the term 'adjusted' at the end of the variable name.

Variable	Coefficient	Variable	Coefficient
intercept	-0.29	(26) Financial and insurance activities	-0.65
Stock	-0.09	(24) Information and communication	0.04
Debtors	0.33	(19) Real estate activities	-0.21
OtherCurrentAssets	-1.30	(29) Electricity, gas, steam and air conditioning supply	-0.21
CashCashEquivalent	1.15	(18) Arts, entertainment and recreation	0.08
Capital	-0.20	(13) Accommodation and food service activities	0.26
OherShareholdersFunds	-1.55	(16) Transportation and storage	-0.35
TotalSharehFundsLiab	-1.60	(21) Other service activities	0.39
CurrentRatio	-0.03	(28) Water supply; sewerage, waste management and remediation activities	0.50
TangibleNetWorth	3.22	StockAdjusted	-0.53
Solvency	-0.02	DebtorsAdjusted	-0.43
TotalSharehFundsLiab-Stock	0.03	OtherCurrentAssetsAdjusted	1.75
TotalSharehFundsLiab-OtherCurrentAssets	-0.22	CashCashEquivalentAdjusted	-0.76
TangibleNetWorth-Debtors	-0.13	CapitalAdjusted	0.26
TangibleNetWorth-Stock	-0.06	OherShareholdersFundsAdjusted	0.76
TangibleNetWorth-TotalSharehFundsLiab	0.01	TotalSharehFundsLiabAdjusted	-1.46
TotalSharehFundsLiab-Debtors	0.00	CurrentRatioAdjusted	0.04
(12) Professional, scientific and technical activities	-0.19	TangibleNetWorthAdjusted	0.88
(27) Mining and quarrying	0.40	SolvencyAdjusted	-0.04
(25) Manufacturing	-0.29	TotalSharehFundsLiab-StockAdjusted	-0.31
(14) Wholesale and retail trade; repair of motor vehicles and motorcycles	-0.18	TotalSharehFundsLiab-OtherCurrentAssetsAdjusted	0.23
(15) Financial and insurance activities	-0.18	TangibleNetWorth-DebtorsAdjusted	0.23
(22) Administrative and support service activities	-0.14	TangibleNetWorth-StockAdjusted	-0.17
(17) Education	-0.19	TangibleNetWorth-TotalSharehFundsLiabAdjusted	-0.22
(23) Human health and social work activities	-0.07	TotalSharehFundsLiab-DebtorsAdjusted	-0.07
(20) Agriculture, forestry and fishing	-0.52		