risks

MDPI

Article

Consumer Bankruptcy Prediction Using Balanced and Imbalanced Data

Magdalena Brygała

Faculty of Management and Economics, Gdansk University of Technology, Narutowicza 11/12, 80-233 Gdansk, Poland; Magdalena.brygala@pg.edu.pl

Abstract: This paper examines the usefulness of logit regression in forecasting the consumer bankruptcy of households using an imbalanced dataset. The research on consumer bankruptcy prediction is of paramount importance as it aims to build statistical models that can identify consumers in a difficult financial situation that may lead to consumer bankruptcy. In the face of the current global pandemic crisis, the future of household finances is uncertain. The change of the macroeconomic and microeconomic situation of households requires searching for better and more precise methods. The research relies on four samples of households: two learning samples (imbalanced and balanced) and two testing samples (imbalanced and balanced) from the Survey of Consumer Finances (SCF) which was conducted in the United States. The results show that the predictive performance of the logit model based on a balanced sample is more effective compared to the one based on an imbalanced sample. Furthermore, mortgage debt to assets ratio, age, being married, having credit constraints, payday loans or payments more than 60 days past due in the last year appear to be predictors of consumer bankruptcy which increase the risk of becoming bankrupt. Moreover, both the ratio of credit card debt to overall debt and owning a house decrease the risk of going bankrupt.

Keywords: bankruptcy of households; prediction; logit; US; household finance; choice-based sample

Citation: Brygała, Magdalena. 2022. Consumer Bankruptcy Prediction Using Balanced and Imbalanced Data. *Risks* 10: 24. https://doi.org/10.3390/risks10020024

Academic Editors: Silvia Dedu and Anatoliy Swishchuk

Received: 22 October 2021
Accepted: 10 January 2022
Published: 18 January 2022

1. Introduction

Personal bankruptcy has grown from a relatively rare household event a couple of decades ago to a fairly common occurrence today (Zhu 2011). Consumer bankruptcy is one of the possibilities that natural persons can use to deal with insolvency that may result from poor financial management, inappropriate consumption habits, unexpected situations, e.g., illness, job loss (Caputo 2008). In 2019 in the United States nearly 751,000 consumers filed for personal bankruptcy (U.S. Courts 2021). Consumer debtors filing for bankruptcy in 2019 reported having total assets of $82 billion and total liabilities of $112 billion (U.S. Courts 2021). However, there is also a large group in society that is very close to the declaration of consumer bankruptcy, although it has not yet announced bankruptcy, which may change due to, for example, a sudden increase in credit card debt or a shock-related event, for instance, loss of employment, illness, divorce or other family problems or mortgage debt (Sullivan et al. 2000). One of the shocks some households must deal with today during the COVID-19 pandemic are shocks to income and wealth which significantly affect a household's future expenses, debt or professional activity (Hanspal et al. 2020). Based on his research, White (1998) concluded that the number of bankruptcies would double if all debtors who have problems with paying their debts went bankrupt. Valaskova et al. (2021) noted that the Covid-19 pandemic will contribute to a worse financial situation for consumers the longer the imposed restrictions last. They also noticed that during the Covid-19 pandemic, the most important aspect affecting financial situation are income, age and sector of occupation. High household indebtedness contributes to the financial instability of households and increases their probability to default on their credit obligations, especially in the event of adverse income shocks (Jappelli et al. 2013). In the case of bankruptcy of

enterprises, the studies conducted so far indicate that bankruptcy does not occur suddenly, but financial problems begin ahead of time (Kliestik et al. 2018), the situation is similar in the case of consumer default (Albanesi and Vamossy 2019).

A crucial issue in financial decisions is the ability to accurately predict consumer bankruptcy or problems with repayment of liabilities. In the literature, various approaches to the topic of bankruptcy risk forecasting can be found, including, among others: identification of methods of creating accurate models, examining the role of variables, analyzing the types of failures that the model is able to predict, or analyzing the sample size, or costs of misclassification (Tian et al. 2015; Crone and Finlay 2012; Du Jardin 2010; Min and Lee 2005; Mossman et al. 1998; Karels and Prakash 1987; Ohlson 1980). Consumer bankruptcy continues to be one of several challenges for banks and other lending institutions in many countries all over the world. To minimize the risk of financial loss on the part of financial institutions, it is important to be able to predict in advance the probability of consumer bankruptcy in particular resulting from the changes in the microeconomic and macroeconomic situation of households. It is also crucial to mention that due to social changes, bankruptcy is generally more acceptable than it was some time ago (Zywicki 2004). Therefore, it may contribute to changes in the number of consumer bankruptcies and changes in bankrupts' profiles in the United States and in other countries around the world.

The focus in this study is on the development of a personal bankruptcy prediction model which is based on two samples: balanced and imbalanced. The study includes a logistic regression function in which, by selecting demographic and financial variables, it is possible to identify households that have financial problems that may lead to bankruptcy. Therefore, the main goal of this study is to develop prognostic models of consumer bankruptcy based on data from the United States. Using data from the Survey of Consumer Finances (SCF), the models for balanced and imbalanced data will be compared and the effectiveness of each model will be determined. Furthermore, the process of setting the optimized cut-off point is employed in this study. Finally, the most appropriate variables for model development will be investigated.

The contribution of this paper is to crucially supplement the existing literature as this study analyzes debtors who filed for consumer bankruptcy in the years 2002–2019 in the United States, which was before the global pandemic crisis. The data includes, among others, the most recent study from the SCF which was conducted in 2019. As most of the research in the literature focuses on predicting corporate bankruptcy as well as nonperforming loans rather than predicting consumer bankruptcy, a logit model for balanced and unsustainable samples was used and compared to the prediction of consumer bankruptcy as exemplified by data from the United States. There are still not enough models for predicting consumer bankruptcy in the literature. Finally, this study has successfully presented consumer bankruptcy profiles and relevant personal bankruptcy variables using demographic variables, ratios related to liabilities or variables such as delays in repayments, payday loans and credit constraints.

The paper is organized into four sections; in the introduction, the author substantiates the topic, research objectives and contribution to the literature. Section 2 explains the data and methods used in the analysis. Section 3 discusses the results, and Section 4 offers some concluding remarks.

2. Literature Review

There are various reasons why debtors file for bankruptcy. What most of them have in common is that households expect immediate benefits from filing for bankruptcy that outweigh the sustainable costs (Evans and Bauchet 2017). The literature showed that the causes of consumer bankruptcy can be divided into macroeconomic and microeconomic factors. Macroeconomic factors take into account interest rates, exchange rates, GDP growth rate, unemployment rate, inflation rate, housing market (Korol 2021a; Bauchet and Evans 2019; Dawsey 2014; Jappelli et al. 2013; Fay et al. 2002). One of the macroeconomic factors influencing consumer bankruptcy is the interest rates. An increase in interest rates on loans

or credit cards increases the monthly burden on households (Ellis 1998). Moreover, Gross and Souleles (2002) pointed out that higher unemployment and lower house prices are associated with more bankruptcies. This is due to the fact that growing unemployment makes borrowers unable to pay off their liabilities. Another important factor is the exchange rate, which can directly affect the household through foreign currency loans or indirectly through the increase in the price of imported goods such as gas.

Among microeconomic aspects, the most common factors include age, education, marital status, gender, homeownership, type and number of debts, income and number of dependent children (Syed Nor et al. 2019; Dawsey 2014; Fisher 2005; Zywicki 2004). In the literature on the subject, attention has been paid to the relationship between marital status and bankruptcy (e.g., Fisher 2019; Agarwal et al. 2011; Fay et al. 2002). A change in marital status can cause deterioration of the consumer's financial situation, e.g., due to the death of a spouse or divorce. Moreover, Fisher and Lyons (2006) noticed that divorce significantly increases the probability of bankruptcy. Although the income factor is often found in the literature as a predictor of consumer bankruptcy (Zhu 2011; Fay et al. 2002), the research carried out by Bauchet and Evans (2019) shows that income was not statistically significantly related to the probability of filing for household bankruptcy. Bauchet and Evans (2019) also pointed to the lack of a statistically significant relationship between filing for personal bankruptcy and self-employment, which was also confirmed in these studies. According to Domowitz and Sartain (1999), owning a house affects both the decision to file for consumer bankruptcy and the choice of a bankruptcy procedure. Several sources in the literature have found such a relationship (e.g., Stavins 2000; Agarwal et al. 2011). The reason can be associated with the risk of losing a home to pay off debts. The literature review also showed that there are several sources that associate debt with filing for consumer bankruptcy. Filing for personal bankruptcy may be connected with credit card debt, medical debt, mortgage debt, car loans or educational loans (Bauchet and Evans 2019; Syed Nor et al. 2019; Zhu 2011; Gross and Souleles 2002; Domowitz and Sartain 1999). Zhu (2011) also reported that bankrupt households took out more credit cards, mortgages and mobile loans with lower average income. Skiba and Tobacman (2019) noted that access to payday loans seems to encourage bankruptcy applications because it deteriorates the financial situation of consumers through their annual interest rates of several hundred percent. Consumers can decide on payday loans despite the high costs because they are not able to obtain a loan from a bank and have been refused or have expected to be refused credit (credit constraints) or because they do not meet the conditions to be granted a loan, e.g., in a bank. Debt repayment behaviors or a consequence of e.g., having too many loans and inadequate management of the household budget, such as late repayment may constitute an early warning of impending bankruptcy (Moorman and Garasky 2008; Himmelstein et al. 2005). Lozinskaia et al. (2016) noted that the accumulation of mortgage loans depending on the level of loan amount and the value of the asset led to a discontinuity in the relative credit loss in the event of mortgage default. Alfaro and Gallardo (2012) also drew attention to the type of defaulted liabilities. They pointed out that the level of education is a factor determining mortgage defaults, while non-payment of consumer liabilities is determined by age and the number of people in the household.

It is worth paying attention to factors related to the approach to spending money, saving, taking loans, using credit cards for daily expenses, compulsive shopping, and expectations about future earnings as they are considered to be linked with debt decisions and consumer bankruptcy (Korol 2021b; Roberts and Jones 2001). Inadequate consumption habits can lead to taking credits, excessive use of credit cards and, as a consequence, liabilities that are too high to maintain financial liquidity.

3. Materials and Methods

3.1. Data

The author used microdata from the SCF. According to the changes in bankruptcy law in 2005, the models contain data collected between 2007 and 2019. During this period,

24,522 surveys were conducted among households. The SCF is a cross-sectional survey conducted in the United States, typically every three years. It covers household information such as demographic, behavioral and financial characteristics. The multiple imputation technique was used to allow for the unanswered questions in the survey. Missing data in the survey have been imputed five times.

In the estimated model, the dependent variable takes value 1 for households who decided to apply for bankruptcy and 0 otherwise. The explanatory variables include the main financial and socio-demographic characteristics of households (Table 1). The considered economic and socio-demographic characteristics were consistent with those commonly used in the literature. The dummy variable year has been included to control for aggregate economic effect. Therefore, the following factors are included in the model and the a priori hypotheses are as follows. Having a house or being male are negatively related to personal bankruptcy (Fay et al. 2002; Fisher 2019; Evans and Bauchet 2017). The income to total debt ratio is expected to be negatively related to bankruptcy (Domowitz and Sartain 1999). Furthermore, having payday loans, being late in repayments, having credit constraints or the credit card debt to the total debt ratio are positively related to personal bankruptcy (Skiba and Tobacman 2019; Korol 2021b). Moreover, age, the number of children or work status are also positively related to personal bankruptcy (Evans and Bauchet 2017; Alfaro and Gallardo 2012). Being married is also expected to be a positive sign (Moorman and Garasky 2008). Thus, a positive relationship is hypothesized between bankruptcy and housing debt to the value of total assets ratio.

Table 1. The list of variables used in evaluating logit models. Source: based on own studies.

Variable	Description
income/debt	It represents the share of housing income in the total debt.
credit card debt/debt	It shows the share of credit card debt in the total debt.
mortgage/assets	It represents the proportion of housing debt to the value of total assets.
late60	The dummy variable of 1 if the household had any payments more than 60 days past due in the last year.
hpayday	The dummy variable of 1 if the household has a payday loan.
education	The variable education is described by four values: 0: no high school, 1: high school, 2: college or associate degree, 3: Bachelor's degree or higher.
house	The dummy variable homeownership class is described by two values: 1: owns e.g., ranch/farm/mobile home/house/condo, 0: otherwise.
married	The dummy variable of 1 if the respondent is married or living with a partner.
male	The dummy variable of 1 if the respondent is male.
age	The variable age is described by six values: 1: <35, 2: 35–44, 3: 45–54, 4: 55–64, 5: 65–74, 6: $\geq$75.
children	The number of children.
work status	The variable work status is described by four values: 0: work for someone else, 1: self-employed/partnership, 2: retired/disabled + student/homemaker, 3: other groups not working.
turndown	The dummy variable of 1 if the respondent applied for a loan in the past 12 months and feared denial or was turned down.
year 2007	The dummy variable of 1 if the survey was from 2007.
year 2010	The dummy variable of 1 if the survey was from 2010.
year 2013	The dummy variable of 1 if the survey was from 2013.
year 2016	The dummy variable of 1 if the survey was from 2016.

Descriptive statistics of selected variables are presented in Table 2, for groups, i.e., bankrupt and non-bankrupt households to demonstrate the basic characteristics of the variables in the sample for balanced and imbalanced datasets. For education, over 21,76%

of households who filed for bankruptcy had bachelor's degrees or higher education, while 9,7% had less than high school education. Among households that did not file for bankruptcy, 34.28% had bachelor's degrees or higher education and 9.4% had not completed high school. Delays in repayment of liabilities exceeding 60 days were observed in 21.18% of people who filed for bankruptcy. Only 8.64% of those who did not file for bankruptcy had delays in payment longer than 60 days. Moreover, looking closer at the demographic characteristics of the learning samples, it can be seen that households that have decided to file for bankruptcy have a lower income to total debt and a higher mortgage debt to asset ratio. Furthermore, 50.59% of households seeking bankruptcy and 65.14% of non-bankrupt households have an owned principal house. Forty percent of bankrupt households who applied for a loan during the last 12 months were refused or did not apply because they expected the loan to be refused. Among non-bankrupt households, only 16.75% of households had such refusals or did not apply for a loan because they expected a refusal.

Table 2. Descriptive statistics of variables selected from the preliminary analysis of the learning data for balanced and imbalanced datasets. Source: based on own research.

	Imbalanced Dataset				**Balanced Dataset**			
	Bankrupt		**Non-Bankrupt**		**Bankrupt**		**Non-Bankrupt**	
	N = 340		**N = 8100**		**N = 340**		**N = 340**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
income/debt	35.219	26.189	43.403	29.926	8.961	1.713	13.732	3.881
credit card debt/debt	0.080	0.012	0.145	0.003	0.070	0.011	0.151	0.017
mortgage/assets	0.241	0.018	0.203	0.003	0.265	0.020	0.207	0.015
education	1.659	0.050	1.880	0.011	1.686	0.049	1.809	0.056
house	0.506	0.027	0.651	0.005	0.535	0.027	0.656	0.026
late60	0.213	0.023	0.086	0.003	0.205	0.022	0.081	0.015
hpayday	0.103	0.017	0.042	0.002	0.103	0.017	0.018	0.007
married	0.641	0.026	0.609	0.005	0.635	0.026	0.612	0.027
male	0.706	0.025	0.749	0.005	0.732	0.024	0.741	0.024
age	1.715	0.065	1.857	0.016	1.738	0.068	1.800	0.081
children	1.185	0.066	0.915	0.013	1.138	0.067	0.888	0.064
work status	0.560	0.050	0.640	0.010	0.580	0.050	0.680	0.05
turndown	0.400	0.027	0.168	0.004	0.377	0.026	0.177	0.021
year 2007	0.168	0.020	0.140	0.004	0.147	0.019	0.147	0.019
year 2010	0.247	0.023	0.228	0.005	0.256	0.024	0.256	0.024
year 2013	0.274	0.024	0.208	0.005	0.277	0.024	0.277	0.024
year 2016	0.200	0.022	0.224	0.005	0.203	0.022	0.203	0.022

3.2. Methodology

Because of utilizing imbalanced data, this study deployed the random down-sample technique, which randomly excludes the observations in the majority to equalize the sample. Predicting rare events (e.g., loan defaults, bankruptcies) is often challenging due to the problem of unsustainable data. The conducted survey includes an imbalanced number of households that have applied for and have not applied for bankruptcy in the last five years. According to Akosa (2017), the results show that imbalanced data can affect the performance of a model (e.g., scorecard, logit model and decision tree models). The techniques which can be used to improve the performance of models and classification downsample the majority class or oversample the minority class (Wah et al. 2016). Imbalanced data can affect the sensitivity of a sample despite high accuracy. Syed Nor et al. (2019) in their studies noticed an improvement in specificity rate applying the random undersampling strategy which improved the classification of bankruptcies of individuals and the prediction of DT model performance. Improving accuracy is one of the most important problems raised in

predicting bankruptcy, the aim of which is to assess the conditions under which the model works well (Du Jardin 2010).

The author created four samples: two learning (imbalanced and balanced) and two testing (imbalanced and balanced) samples. The imbalanced dataset of both learning and testing samples included 8,440 consumers who have any debt. The balanced dataset of both learning and testing samples included 340 consumers who have any debt. The division of the learning sample and the testing sample was created to enable the estimation of prognostic models on a learning sample and then their testing on an unknown testing sample. This approach is used in the literature on the subject (Syed Nor et al. 2019; Irimia-Dieguez et al. 2015; Chen 2011). For each model, observations from the years 2007–2019 have been pooled.

One of the most widely used credit scoring techniques is logistic regression (Abdou and Pointon 2011; Lee et al. 2002; Laitinen 1999; Westgaard and Van der Wijst 2001). Credit scoring models are methods commonly used for predicting personal bankruptcy (Xiong et al. 2013). The role of credit scoring is to support and help financial institutions, especially banks, in maximizing the expected profit from a client by reducing the probability of default by a client (Abdou and Pointon 2011). Several studies have shown that logistic models can be applied to predict bankruptcy (e.g., Korol 2021a; Bateni and Asghari 2020; Son et al. 2019; Mihalovic 2016; Irimia-Dieguez et al. 2015; Chen 2011; Back et al. 1996). In the estimation of the probability of becoming bankrupt, the logit model has been applied. The logistic regression equation can be written as an odds ratio:

$$\frac{\pi}{1-\pi} = \exp(\alpha + \beta_1 X_1 + \cdots + \beta_M X_M), \tag{1}$$

and it can be converted to the following form (Peng et al. 2002):

$$logit(Y) = \ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 X_1 + \cdots + \beta_M X_M, \tag{2}$$

where π is the probability of the event, α is the Y intercept, β's are regression coefficients and X's are predictors. The model coefficients are estimated by maximum-likelihood from the dataset.

The performance measure consists of the following indicators: total effectiveness (S), type I error (E_1), type II error (E_2). The total effectiveness of a prediction model shows the probability of a correct prediction of bankruptcy and non-bankruptcy. The indicator shows the overall performance of the model and it is estimated as (Korol 2021b):

$$S = (1 - \frac{D_1 + D_2}{BR + NBR}) * 100\%, \tag{3}$$

where type I error shows false predictions of bankrupts (D_1) to all bankrupts (BR) and type II error–false prediction of non-bankrupts (D_1) to all non-bankrupts (NBR). Type I error is calculated as:

$$E_1 = \frac{D_1}{BR} * 100\%. \tag{4}$$

Type II error is computed as:

$$E_2 = \frac{D_2}{NBR} * 100\%. \tag{5}$$

After estimating the logit model, it is essential to find the optimal cut-off point to correctly classify households into the groups of bankrupt and non-bankrupt households. The correct selection of the cut-off point determines the classification results (Mihalovic 2016). Looking only at the overall efficiency in the case of unbalanced data may lead to the distortion of the results and a negligible classification of a smaller group, e.g., bankrupts, compared to the larger group of non-bankrupts. The optimal cut-off point minimizes type I error

and type II error and maximizes sensitivity and specificity (Affes and Hentati-Kaffel 2019). Type I error is a situation in which the consumer is not classified as a potential bankrupt and will have problems with repayment of liabilities, which will lead to bankruptcy. Granting a loan to a person who will have problems with repayment of liabilities and, consequently, declares bankruptcy may lead to losses, for example on the part of a bank, due to the inability to recover the loan. Type II error is a loss of a potential customer and profit for a bank or another lending institution in connection with not granting a loan to a person who would have no problems with repayment of liabilities and would not declare bankruptcy. Consequently, type I error is considered to be more expensive than type II error (West 2000). The costs of misclassification should also be considered depending on the subject of research, because the costs of misclassification in medicine will be considered differently than, for example, in banking. Minimization of both type I error and type II error guarantees the lending institution a low risk of default of borrowers and the maximization of loans granted due to consumer bankruptcy. The commonly used threshold is c = 0.5 (Couronné et al. 2018). Moreover, because of assuming the cut-off point at the level of 0.5, we assume that the loss function is symmetric for the two error rates, so choosing a cut-off point at this level should not be a standard choice as it may not always be the most appropriate (Ohlson 1980).

4. Results and Discussion

Table 3 shows the significant variables of the logit model for the balanced and imbalanced data. There are 11 significant variables for the imbalanced dataset and 7 for the balanced dataset. Models using imbalanced and balanced data show that owning a house reduces the likelihood of filing for bankruptcy and it is one of its most important predictors, which is in agreement with researches carried out by e.g., Syed Nor et al. (2019), Fisher (2005). Another strong determinant of filing for bankruptcy is the housing debt to asset ratio. A higher rate contributes to a higher probability of filing for personal bankruptcy. Qi and Yang (2009) analyzed the loss on default and concluded that the loan-to-value mortgages ratio is an important determinant. In my research, the mortgage debt to total assets ratio was analyzed but the results also indicate the significance of this variable.

Table 3. The significant variables of the logit models for the balanced and imbalanced datasets. Source: based on own research.

Imbalanced Data	Balanced Data
credit card debt/debt	credit card debt/debt
mortgage/asset	mortgage/asset
house	house
late60	late60
male	hpayday
married	age
age	turndown
turndown	
year 2007	
year 2010	
year 2013	

Households with credit constraints (those who reported being denied credit in the past year, as well as those who did not apply for credit in the past years due to fear of being denied), or with any payments more than 60 days past due in the last year have higher bankruptcy risk. However, not every consumer who is delayed will decide to file for consumer bankruptcy. The duration of delays could also be crucial. Some of the studies on personal bankruptcy analyzing models of consumer bankruptcy do not take into account delays in repayment as one of the determinants leading to bankruptcy. Only a few studies have analyzed consumer behavior in terms of debt payment as one of the determinants of bankruptcy. My results support research showing that debt repayment behavior such

as having delays in repayment during the last period may be one of the bankruptcy determinants (Moorman and Garasky 2008; Himmelstein et al. 2005). Consumer behavior in terms of debt repayment can also contribute to credit constraints both on the part of the lender and resulting from the consumer's fear of refusal to being granted a loan. Having credit constraints is a strong determinant of filing for bankruptcy. Refusal to grant a loan or fear of applying for a loan (which is not always grounded and may result from a lack of financial knowledge) may aggravate the household's financial problems and lead to consumer bankruptcy.

Credit cards are considered a strong determinant of bankruptcy (Zhu 2011; Gross and Souleles 2002), but bankruptcy is also influenced by the share of various liabilities, which is also important. Having a higher ratio of credit card debt to all debt reduces the likelihood of filing for bankruptcy. Furthermore, the number of children does not statistically significantly affect the probability of applying for consumer bankruptcy. Similar conclusions can be found in the research conducted by Moorman and Garasky (2008), they showed that family size has no statistically significant relationship with filing for bankruptcy. Moreover, age is also a strong determinant of filing for bankruptcy, and it changes with age group. The highest probability of applying for consumer bankruptcy is noticeable in the age group 45–54 and 55–64 compared to people under 35 years of age. The results are in agreement with Bauchet and Evans (2019) who concluded that age increases the likelihood of filing for bankruptcy. They also pointed out that the relationship between age and bankruptcy was non-linear, so the probability of bankruptcy increased at a decreasing rate with age.

Of note, research has shown that employment status has no statistically significant relationship with applying for consumer bankruptcy, which was also confirmed by Bauchet and Evans (2019). However, these results are in disagreement with Zhu's (2011). Models using imbalanced data show that being married/in a relationship or male increases the likelihood of filing for bankruptcy. Moorman and Garasky (2008) also noted that being married or male increases the likelihood of filing for bankruptcy. Bauchet and Evans (2019) came to different conclusions that being married contributes to a lower possibility of filing for bankruptcy. Furthermore, the model using balanced data shows that having a payday loan increases the likelihood of filing for bankruptcy. Skiba and Tobacman (2019), Martin and Tong (2009) got similar results from their research that having payday loans contributes to filing for bankruptcy according to the worsening financial situation of the households.

The results including regression coefficients and standard errors for the logit models are presented in Table 4. The estimations are run using the imbalanced sample (8440 records) and the balanced sample (680 records). Results are displayed for the learning sample.

Table 5 presents the prediction results for the logit model along with the cut-off points determined for imbalanced data. There is a clear difference in studies between balanced and imbalanced data. In the case of imbalanced data, the cut-off point of 0.5 in the learning sample gives us a high efficiency of 95.96%, but with a high type I error at 100% and a low type II error at 0.01%. The situation is similar in the case of testing data. Using a cut-off point of 0.5 yields a total efficiency of 95.98% but with a type I error—of 99.71% and the type II error—of 0%. Out of 340 households that filed for bankruptcy, only one was correctly identified as bankrupt for testing data and zero for learning data. Therefore, such a model is practically useless, and the high-efficiency results only from the fact that there was a disproportion between bankrupts and non-bankrupts. It follows that for virtually all consumers the model recognizes that there is no risk of bankruptcy.

Table 4. Results of logit models. Source: based on own research.

Variables		Model Imbalanced Coefficients (B)	Model Imbalanced S.E.	Model Balanced Coefficients (B)	Model Balanced S.E.	Base Unit
income/debt		0.000	0.000	0.001	0.001	
credit card debt/debt		−1.023 **	0.311	−1.018 *	0.446	
mortgage/assets		1.208 ***	0.282	1.595 **	0.507	
education	high school	0.291	0.220	0.432	0.327	Less than high school education
	college or associate degree	0.042	0.223	0.174	0.332	
	bachelor's degree or higher	−0.218	0.244	−0.121	0.346	
house		−1.223 ***	0.215	−1.234 ***	0.301	
late60		0.468 **	0.174	0.972 **	0.319	
hpayday		0.250	0.223	1.468 **	0.499	
married		0.718 ***	0.204	0.462	0.335	
male		−0.658 **	0.213	−0.440	0.350	
age	age: 35−44	0.669 ***	0.197	0.508	0.281	age: <35
	age: 45−54	1.024 ***	0.197	1.181 ***	0.274	
	age: 55−64	0.935 ***	0.224	1.147 ***	0.316	
	age: 65−74	0.602	0.325	0.984*	0.432	
	age: ≥75	0.598	0.488	0.872	0.666	
children		0.006	0.052	0.069	0.087	
work status	work for someone else	0.065	0.262	−0.293	0.389	unemployed
	self−employed/partnership	0.155	0.315	−0.204	0.483	
	retired/disabled + student/homemaker	−0.050	0.309	−0.475	0.485	
turndown		0.825 ***	0.140	0.942 ***	0.226	
year 2007		0.714 **	0.233	0.040	0.343	
year 2010		0.442 *	0.222	−0.386	0.320	
year 2013		0.672 **	0.218	−0.052	0.309	
year 2016		0.415	0.225	−0.015	0.324	
_cons		−4.110 ***	0.394	−0.468	0.588	

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 5. The results of the effectiveness of models together with cut-off points for the imbalanced sample. Source: based on own research.

Training Dataset									
cut-off point	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
Type I error	0.88%	8.53%	21.47%	31.18%	41.76%	49.41%	54.41%	61.76%	67.65%
Type II error	90.58%	66.86%	45.02%	31.00%	23.23%	17.58%	13.78%	10.65%	8.58%
Total effectiveness	13.03%	35.49%	55.92%	68.99%	76.02%	81.14%	84.59%	87.29%	89.04%
cut-off point	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Type I error	72.94%	95.88%	99.41%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Type II error	6.86%	0.96%	0.12%	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%
Total effectiveness	90.47%	95.21%	95.88%	95.96%	95.96%	95.97%	95.97%	95.97%	95.97%
Testing Dataset									
cut-off point	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
Type I error	2.65%	12.06%	23.82%	37.65%	45.88%	52.06%	57.06%	63.24%	70.00%
Type II error	90.73%	67.37%	44.95%	31.53%	23.60%	18.10%	14.04%	11.01%	8.79%
Total effectiveness	12.82%	34.86%	55.90%	68.22%	75.50%	80.53%	84.23%	86.88%	88.74%
cut-off point	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Type I error	74.12%	96.18%	99.12%	99.71%	99.71%	99.71%	99.71%	100.00%	100.00%
Type II error	6.99%	1.09%	0.07%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%
Total effectiveness	90.31%	95.08%	95.94%	95.97%	95.98%	95.98%	95.98%	95.97%	95.97%

For imbalanced data, predictions show that 0.04 seems to be the optimal cut-off point for both the learning dataset and the testing dataset. The total error rate is at the level of 31.01% for the learning dataset and 31.78% for the testing dataset. The total effectiveness of

bankrupt households moves in the opposite direction to that of non-bankrupt households and therefore reducing the occurrence of one type of error leads to increasing the other type of error. It is worth noting that, despite its lower effectiveness, this model is useful because it actually distinguishes bankrupts from non-bankrupts.

Table 6 presents the prediction results for the logit model along with the determined cut-off points for balanced data. Both for learning and testing samples, the optimal cut-off point is at the level of 0.5. For that cut-off point, the total efficiency for testing data is 69.85%, with 29.41% for type I error and 30.88% for type II error. The overall prediction efficiency of models has changed from 68.22% in unbalanced data to 69.85% in balanced data. Type I error decreased from 37.65% for unbalanced data to 29.41% for balanced data. Moreover, type II error decreased from 31.53% for unbalanced data to 30.88% for balanced data.

Table 6. The results of the effectiveness of models together with cut-off points for balanced data. Source: based on own research.

Training Dataset									
cut-off point	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
Type I error	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Type II error	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	99.41%
Total effectiveness	50.00%	50.00%	50.00%	50.00%	50.00%	50.00%	50.00%	50.00%	50.29%
cut-off point	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Type I error	0.00%	2.06%	6.18%	19.12%	32.94%	45.59%	64.71%	82.35%	93.82%
Type II error	98.82%	91.18%	67.65%	46.47%	30.00%	16.76%	8.82%	4.41%	0.59%
Total effectiveness	50.59%	53.38%	63.09%	67.21%	68.53%	68.82%	63.24%	56.62%	52.79%
Testing Dataset									
cut-off point	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
Type I error	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Type II error	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	99.71%	99.41%	99.12%
Total effectiveness	50.00%	50.00%	50.00%	50.00%	50.00%	50.00%	50.15%	50.29%	50.44%
cut-off point	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Type I error	0.00%	2.35%	7.35%	17.94%	29.41%	44.12%	61.18%	75.88%	88.53%
Type II error	98.82%	90.88%	73.53%	50.29%	30.88%	19.71%	10.88%	5.59%	2.06%
Total effectiveness	50.59%	53.38%	59.56%	65.88%	69.85%	68.09%	63.97%	59.26%	54.71%

Studies have shown that it is better to use the balanced sample in connection with higher efficiency and lower type I error and type II error compared to the imbalanced sample, which is confirmed by other studies such as Zhou's (2013) and García et al. (2012). Syed Syed Nor et al. (2019) predicted personal bankruptcy through a decision tree for a balanced and unbalanced sample. The model in the balanced sample showed lower efficiency but also a lower type I error and, therefore, it predicts bankruptcies more effectively, which is confirmed in my research.

The efficiency of the imbalanced sample is not much lower than that of balanced samples, but only while maintaining the optimal cut-off point. Given the same cut-off points for the balanced and imbalanced samples at the level of 0.5, the model for the imbalanced sample is practically useless as it is ineffective in predicting bankruptcy. Moreover, using the optimal cut-off point for the imbalanced sample yields a higher type I error (37.65%) than using it for the balanced sample (29.41%), which is important as type I error is considered more costly for the lender. These results are in agreement with the research carried out by e.g., Mihalovic (2016); Zhou and Elhag (2007), and Chi and Tang (2006), which pay attention to the importance of choosing the optimal cut-off point due to the crucial impact on the predictive results of the business bankruptcy models and credit risk models. My research shows the effect of the cut-off point on the effectiveness of household bankruptcy models.

5. Conclusions

Consumer bankruptcy is still a very important issue due to the changing microeconomic and macroeconomic situation of households. Predicting rare events such as consumer bankruptcy is often difficult due to the problem of unsustainable data and may cause bias in the estimated bankruptcy probabilities. This article discusses and compares bankruptcy classification using random undersampling to correct unsustainable data. The use of the random undersampling technique in the logit model showed that the total performance increased, and error rates decreased after using the random subsampling strategy.

In summary, the applied research approach provided clear evidence that the predictive performance of the logit models based on a balanced sample is more effective compared to those based on an imbalanced sample. These models showed fewer type I errors than type II errors and yielded the highest overall effectiveness of forecast. It is also worth noting that if we used the same cut-off points for a balanced and imbalanced sample at the level of 0.5, the model for an imbalanced sample is practically useless and it is ineffective in predicting bankruptcy. Despite the high total effectiveness of such a model, this is due to the fact that almost all bankrupts are recognized by the model as non-bankrupts. The efficiency of the imbalanced sample is not much lower than that of balanced samples but only when the optimal cut-off points are used. However, using optimal cut-off points for the imbalanced sample yields a higher type I error (37.65%) than using them for the balanced sample (29.41%) which is important as type I error is considered more costly for the lender.

Furthermore, in the presented empirical study it was possible to identify important factors influencing the likelihood of filing for bankruptcy. The mortgage debt to assets ratio, being married, age, having credit constraints, payday loans or payments more than 60 days past due in the last year increase the risk of becoming bankrupt. Moreover, the credit card debt to overall debt ratio, and owning a house decrease the risk of becoming bankrupt.

The author is aware of various limitations of the conducted study. Its main limitation is limited data access. In future research, the author will continue research into the use of various techniques that can be employed to deal with unbalanced datasets, using oversampling, undersampling, bagging or boosting methods to improve the performance of the logit model, decision trees, multivariate discriminant analysis, random forest or fuzzy logic.

Funding: This research received no external funding.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found at: https://www.federalreserve.gov/econres/scfindex.htm (accessed on 15 April 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

Abdou, Hussein A., and John Pointon. 2011. Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting, Finance and Management* 18: 59–88. [CrossRef]

Affes, Zeineb, and Rania Hentati-Kaffel. 2019. Predicting US banks bankruptcy: Logit versus Canonical Discriminant analysis. *Computational Economics* 54: 199–244. [CrossRef]

Agarwal, Sumit, Souphala Chomsisengphet, and Chunlin Liu. 2011. Consumer bankruptcy and default: The role of individual social capital. *Journal of Economic Psychology* 32: 632–50. [CrossRef]

Akosa, Josephine. 2017. Predictive accuracy: A misleading performance measure for highly imbalanced data. Paper presented at SAS Global Forum, Orlando, FL, USA, April 2–5; vol. 12.

Albanesi, Stefania, and Domonkos F. Vamossy. 2019. *Predicting Consumer Default: A Deep Learning Approach*; No. w26165; Cambridge: National Bureau of Economic Research.

Alfaro, Rodrigo, and Natalia Gallardo. 2012. The determinants of household debt default. *Revista de Analisis Economico* 27. [CrossRef]

Back, Barbro, Teija Laitinen, Kaisa Sere, and Michiel van Wezel. 1996. Choosing bankruptcy predictors using discriminant analysis, logit analysis, and genetic algorithms. *Turku Centre for Computer Science Technical Report* 40: 1–18.

Bateni, Leila, and Farshid Asghari. 2020. Bankruptcy prediction using logit and genetic algorithm models: A comparative analysis. *Computational Economics* 55: 335–48. [CrossRef]

Bauchet, Jonathan, and David Evans. 2019. Personal bankruptcy determinants among US households during the peak of the Great Recession. *Journal of Family and Economic Issues* 40: 577–91. [CrossRef]
Caputo, Richard K. 2008. Marital status and other correlates of personal bankruptcy, 1986–2004. *Marriage & Family Review* 44: 5–32.
Chen, Mu-Yen. 2011. Predicting corporate financial distress based on integration of decision tree classification and logistic regression. *Expert Systems with Applications* 38: 11261–72. [CrossRef]
Chi, Li-Chiu, and Tseng-Chung Tang. 2006. Bankruptcy prediction: Application of logit analysis in export credit risks. *Australian Journal of Management* 31: 17–27. [CrossRef]
Couronné, Raphael, Philipp Probst, and Anne-Laure Boulesteix. 2018. Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics* 19: 1–14. [CrossRef]
Crone, Sven F., and Steven Finlay. 2012. Instance sampling in credit scoring: An empirical study of sample size and balancing. *International Journal of Forecasting* 28: 224–38. [CrossRef]
Dawsey, Amanda E. 2014. Externalities among creditors and personal bankruptcy. *Journal of Financial Economic Policy* 6: 2–24. [CrossRef]
Domowitz, Ian, and Robert L. Sartain. 1999. Determinants of the consumer bankruptcy decision. *The Journal of Finance* 54: 403–20. [CrossRef]
Du Jardin, Philippe. 2010. Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy. *Neurocomputing* 73: 2047–60. [CrossRef]
Ellis, Diane. 1998. *The Effect of Consumer Interest Rate Deregulation on Credit Card Volumes, Charge-Offs, and the Personal Bankruptcy Rate*; FDIC Division of Insurance Paper 98-05; Washington, DC: Federal Deposit Insurance Corporation. Available online: https://www.fdic.gov/bank/analytical/bank-trends/bt9805.pdf (accessed on 10 May 2021).
Evans, David, and Jonathan Bauchet. 2017. Bankruptcy determinants among US households during the peak of the great recession. *Consumer Interests Annual* 63: 1–7.
Fay, Scott, Erik Hurst, and Michelle J. White. 2002. The household bankruptcy decision. *American Economic Review* 92: 706–18. [CrossRef]
Fisher, Jonathan D. 2005. The effect of unemployment benefits, welfare benefits, and other income on personal bankruptcy. *Contemporary Economic Policy* 23: 483–92. [CrossRef]
Fisher, Jonathan D. 2019. Who files for personal bankruptcy in the United States? *Journal of Consumer Affairs* 53: 2003–26. [CrossRef]
Fisher, Jonathan D., and Angela C. Lyons. 2006. Till debt do us part: A model of divorce and personal bankruptcy. *Review of Economics of the Household* 4: 35–52. [CrossRef]
García, Vicente, José Salvador Sánchez, and Ramón Alberto Mollineda. 2012. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems* 25: 13–21. [CrossRef]
Gross, David B., and Nicholas S. Souleles. 2002. An empirical analysis of personal bankruptcy and delinquency. *The Review of Financial Studies* 15: 319–47. [CrossRef]
Hanspal, Tobin, Annika Weber, and Johannes Wohlfart. 2020. *Income and Wealth Shocks and Expectations during the COVID-19 Pandemic*. CEBI Working Paper No. 13/20. Munich: Center for Economic Studies and ifo Institute (CESifo).
Himmelstein, David U., Elizabeth Warren, Deborah Thorne, and Steffie Woolhandler. 2005. Illness and injury as contributors to bankruptcy. *Health Affairs* 24: 570. [CrossRef] [PubMed]
Irimia-Dieguez, Ana Isabel, A. Blanco-Oliver, and María José Vazquez-Cueto. 2015. A comparison of classification/regression trees and logistic regression in failure models. *Procedia Economics and Finance* 23: 9–14. [CrossRef]
Jappelli, Tullio, Marco Pagano, and Marco Di Maggio. 2013. Households' indebtedness and financial fragility. *Journal of Financial Management, Markets and Institutions* 1: 23–46.
Karels, Gordon V., and Arun J. Prakash. 1987. Multivariate normality and forecasting of business bankruptcy. *Journal of Business Finance & Accounting* 14: 573–93.
Kliestik, Tomas, Maria Misankova, Katarina Valaskova, and Lucia Svabova. 2018. Bankruptcy prevention: New effort to reflect on legal and social changes. *Science and Engineering Ethics* 24: 791–803. [CrossRef] [PubMed]
Korol, Tomasz. 2021a. Evaluation of the Macro-and Micro-Economic Factors Affecting the Financial Energy of Households. *Energies* 14: 3512. [CrossRef]
Korol, Tomasz. 2021b. Examining Statistical Methods in Forecasting Financial Energy of Households in Poland and Taiwan. *Energies* 14: 1821. [CrossRef]
Laitinen, Erkki K. 1999. Predicting a corporate credit analyst's risk estimate by logistic and linear models. *International Review of Financial Analysis* 8: 97–121. [CrossRef]
Lee, Tian-Shyug, Chih-Chou Chiu, Chi-Jie Lu, and I-Fei Chen. 2002. Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications* 23: 245–54. [CrossRef]
Lozinskaia, Agata, Evgeniy Ozhegov, and Alexander Karminsky. 2016. *Discontinuity in Relative Credit Losses: Evidence from Defaults on Government-Insured Residential Mortgages*. Working Papers Series: Financial Economics WP BRP 55/FE/2016; Moscow: National Research University Higher School of Economics.
Martin, Nathalie, and Koo Im Tong. 2009. Double down-and-out: The connection between payday loans and bankruptcy. *Southwestern University Law Journal* 39: 785.
Mihalovic, Matús. 2016. Performance comparison of multiple discriminant analysis and logit models in bankruptcy prediction. *Economics & Sociology* 9: 101.

Min, Jae H., and Young-Chan Lee. 2005. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications* 28: 603–14. [CrossRef]
Moorman, Diann C., and Steven Garasky. 2008. Consumer debt repayment behavior as a precursor to bankruptcy. *Journal of Family and Economic Issues* 29: 219–33. [CrossRef]
Mossman, Charles E., Geoffrey G. Bell, L. Mick Swartz, and Harry Turtle. 1998. An empirical comparison of bankruptcy models. *Financial Review* 33: 35–54. [CrossRef]
Ohlson, James A. 1980. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research* 18: 109–31. [CrossRef]
Peng, Chao-Ying Joanne, Kuk Lida Lee, and Gary M. Ingersoll. 2002. An introduction to logistic regression analysis and reporting. *The Journal of Educational Research* 96: 3–14. [CrossRef]
Qi, Min, and Xiaolong Yang. 2009. Loss given default of high loan-to-value residential mortgages. *Journal of Banking & Finance* 33: 788–99.
Roberts, James A., and Eli Jones. 2001. Money attitudes, credit card use, and compulsive buying among American college students. *Journal of Consumer Affairs* 35: 213–40. [CrossRef]
Skiba, Paige Marta, and Jeremy Tobacman. 2019. Do payday loans cause bankruptcy? *The Journal of Law and Economics* 62: 485–519. [CrossRef]
Son, Hwijae, C. Hyun, Du Phan, and Hyung Ju Hwang. 2019. Data analytic approach for bankruptcy prediction. *Expert Systems with Applications* 138: 112816. [CrossRef]
Stavins, Joanna. 2000. Credit card borrowing, delinquency, and personal bankruptcy. *New England Economic Review*, 15–30.
Sullivan, Teresa A., Elizabeth Warren, and Jay Lawrence Westbrook. 2000. *The Fragile Middle Class: Americans in Debt*. New Haven: Yale University Press, vol. 79.
Syed Nor, Sharifah Heryati, Shafinar Ismail, and Bee Wah Yap. 2019. Personal bankruptcy prediction using decision tree model. *Journal of Economics, Finance and Administrative Science* 24: 157–70. [CrossRef]
Tian, Shaonan, Yan Yu, and Hui Guo. 2015. Variable selection and corporate bankruptcy forecasts. *Journal of Banking & Finance* 52: 89–100.
U.S. Courts. 2021. Available online: https://www.uscourts.gov/statistics-reports/caseload-statistics-data-tables (accessed on 10 May 2021).
Valaskova, Katarina, Pavol Durana, and Peter Adamko. 2021. Changes in consumers' purchase patterns as a consequence of the COVID-19 pandemic. *Mathematics* 9: 1788. [CrossRef]
Wah, Yap Bee, HezlinAryani Abd Rahman, Haibo He, and Awang Bulgiba. 2016. Handling imbalanced dataset using SVM and k-NN approach. In *AIP Conference Proceedings*. Melville, NY: AIP Publishing LLC, vol. 1750, p. 020023.
West, David. 2000. Neural network credit scoring models. *Computers & Operations Research* 27: 1131–52.
Westgaard, Sjur, and Nico Van der Wijst. 2001. Default probabilities in a corporate bank portfolio: A logistic model approach. *European Journal of Operational Research* 135: 338–49. [CrossRef]
White, Michelle J. 1998. Why don't more households file for bankruptcy? *Journal of Law, Economics, & Organization* 14: 205–31.
Xiong, Tengke, Shengrui Wang, André Mayers, and Ernest Monga. 2013. Personal bankruptcy prediction by mining credit card data. *Expert Systems with Applications* 40: 665–76. [CrossRef]
Zhou, Ligang. 2013. Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods. *Knowledge-Based Systems* 41: 16–25. [CrossRef]
Zhou, Ying, and Taha M. S. Elhag. 2007. Apply logit analysis in bankruptcy prediction. Paper presented at 7th WSEAS International Conference on Simulation, Modelling and Optimization, Beijing China, September 15–17; pp. 302–8.
Zhu, Ning. 2011. Household consumption and personal bankruptcy. *The Journal of Legal Studies* 40: 1–37. [CrossRef]
Zywicki, Todd J. 2004. An economic analysis of the consumer bankruptcy crisis. *Northwestern University Law Review* 99: 1463. [CrossRef]