APPLICATION OF BIG DATA TECHNOLOGIES IN FINANCIAL/CREDIT RISK ASSESSMENT

Big Data Technologies

Adnan Branković

ABSTRACT

The leader in the digital transformation of the economy these days is the modern technology for processing and analyzing large amount of the data. The reason for this is because of accumulated a huge number of information can not be processed using traditional tools and methods of conventional databases, only with using modern techonolgies that have developed in recent years. Every day in the world, about 20 petabytes of information are generated, and the amount of information is expected to be more than 163 zettabytes by 2025. Experts from the banking and financial sector say that the return on investment in big data solutions helps to improve their personnel management, attract new customers, optimize processes and identify risks exceeds expectations. This helps to low costs, increase productivity, improve customer relationships, predict risks and comply everything with legal requirements and reason for that is because big data technologies provide a higher level of automation. Financial data and related data of enterprises and economy sector provides a massive source of data for early warning of risks arising. Big data analysis technology can realize the application of these massive data to the field of credit risk, which improves the accuracy and scientificity of risk prediction and early warning. Conventional risk assessment methods include Logistic regression method, which can accurately assess credit risk.

INTRODUCTION

Credit risk is the possibility of a loss resulting from a borrower's failure to repay a loan or meet contractual obligations. Traditionally, it refers to the risk that a lender may not receive the owed principal and interest, which results in an interruption of cash flows and increased costs for collection. When lenders offer mortgages, credit cards, or other types of loans, there is a risk that the borrower may not repay the loan. Similarly, if a company offers credit to a customer, there is a risk that the customer may not pay their invoices. Credit risk also describes the risk that a bond issuer may fail to make payment when requested or that an insurance company will be unable to pay a claim.

Credit risks are calculated based on the borrower's overall ability to repay a loan according to its original terms. To assess credit risk on a consumer loan,

lenders look at the five C: credit history, capacity to repay, capital, the loan's conditions, and associated collateral.

The following are three main types of credit risks:

1. **Credit default risk**

Credit default risk occurs when the borrower is unable to pay the loan obligation in full or when the borrower is already 90 days past the due date of the loan repayment. The credit default risk may affect all credit-sensitive financial transactions such as loans, bonds, securities, and derivatives. The level of default risk can change due to a broader economic change. It can also be due because of a change in a borrower's economic situation, such as increased competition or recession, which can affect the company's ability to set aside principal and interest payments on the loan.

2. **Concentration risk**

Concentration risk is the level of risk that arises from exposure to a single counterparty or sector, and it offers the potential to produce large amounts of losses that may threaten the lender's core operations. The risk results from the observation that more concentrated portfolios lack diversification, and therefore, the returns on the underlying assets are more correlated. For example, a corporate borrower who relies on one major buyer for its main products has a high level of concentration risk and has the potential to incur a large amount of losses if the main buyer stops buying their products.

3. **Country risk**

Country risk is the risk that occurs when a country freezes foreign currency payments obligations, resulting in a default on its obligations. The risk is associated with the country's political instability and macroeconomic performance, which may adversely affect the value of its assets or operating profits. The changes in the business environment will affect all companies operating within a particular country.

In order to minimize the level of credit risk, lenders should forecast credit risk with greater accuracy. Listed below are some of the factors that lenders should consider when assessing the level of credit risk:

1. **Probability of Default (POD)**

The probability of default, sometimes abbreviated as POD, is the likelihood that a borrower will default on their loan obligations. For individual borrowers, POD is based on a combination of two factors, i.e., credit score and debt-to-income ratio. The POD for corporate borrowers is obtained from credit rating agencies. If the lender determines that a potential borrower demonstrates a lower probability of default, the loan will come with a low interest rate and low or no down payment on the loan. The risk is partly managed by pledging collateral against the loan.

2. **Loss Given Default (LGD)**

Loss given default (LGD) refers to the amount of loss that a lender will suffer in case a borrower defaults on the loan. For example, assume that two

borrowers, A and B, with the same debt-to-income ratio and an identical credit score. Borrower A takes a loan of $10,000 while B takes a loan of $200,000. The two borrowers present with different credit profiles, and the lender stands to suffer a greater loss when Borrower B defaults since the latter owes a larger amount. Although there is no standard practice of calculating LGD, lenders consider an entire portfolio of loans to determine the total exposure to loss.

3. Exposure at Default (EAD)

Exposure at Default (EAD) evaluates the amount of loss exposure that a lender is exposed to at any particular time, and it is an indicator of the risk appetite of the lender. EAD is an important concept that references both individual and corporate borrowers. It is calculated by multiplying each loan obligation by a specific percentage that is adjusted based on the particulars of the loan.

BIG DATA TECHNOLOGIES

Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software. Data with many fields (columns) offer greater statistical power, while data with higher complexity (more attributes or columns) may lead to a higher false discovery rate. Big data analysis challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy, and data source. Big data was originally associated with three key concepts: volume, variety, and velocity. The analysis of big data presents challenges in sampling, and thus previously allowing for only observations and sampling. Therefore, big data often includes data with sizes that exceed the capacity of traditional software to process within an acceptable time and value.

The growing maturity of the concept more starkly delineates the difference between "big data" and "business intelligence":

- Business intelligence uses applied mathematics tools and descriptive statistics with data with high information density to measure things, detect trends, etc.
- Big data uses mathematical analysis, optimization, inductive statistics, and concepts from nonlinear system identificationto infer laws (regressions, nonlinear relationships, and causal effects) from large sets of data with low information density to reveal relationships and dependencies, or to perform predictions of outcomes and behaviors.

DATA

The dataset, used for the purpose of this research paper, is downloaded from the Kaggle website. Dataset is open-sourced and publicly available. The data quality was assessed under the following headings:

• Accuracy

• Completeness

• Consistency

We can be confirmed that dataset is specifically intended for purpose and meet the above criteria. The columns of dataset are as follows:

1. ID – The unique identifier

2. DEFAULT_IND – A binary indicator that indicates whether the loan is in default or not

3. PORTFOLIO – An indicator whether the loan is a vehicle loan or a mortgage

4. LTV – the ratio of the loan amount to the value of the underlying asset

5. INSTLMNT_AMT – The monthly instalment paid by the obligor

6. AGE – the Age in years of the customer

7. EXT_SCORE – A normalised value indicating the External credit score of the obligor.

The main tool used to perform the analysis was Hadoop. Hadoop provides a software framework for distributed storage and processing of big data using the map-reduce programming model. Map-reduce allows for efficient parallel processing. This type of parallel processing allows for quick and efficient processing of large datasets. The project was developed using a localized Linux virtual machine. This was a user-friendly development environment where various IDEs were utilized.

The cleaned dataset, using Python and cleaning all missing values, was saved as a CSV and then pushed to the Hadoop Distributed File System (HDFS). From there the map-reduce operations to group the data and train a logistic regression.

ANALYSIS

The aim was to find out if a standard credit scorecard model could be built using the map reduce programming model. This involves two steps:

1. Binning each variable and calculating its default rate and its Weight of Evidence (WoE).
2. Training a logistic regression on the calculated weights of evidence.

The weight of evidence (WoE) is a measure of the separation of good and bad customers.

The weight of evidence is calculated for each bin. The right binning chosen is subjective and no single method exists that is widely used across all banks.

For the purposes of this research paper, the decision was taken to use an iterative process for binning. This meant that arbitrary bins were tried until suitable bins that explain the data sufficiently were found.

These groups were summarized using a map-reduce programming structure. The structure of which is as follows:

The input data is a single table from the HDFS. The total values are calculated such as; total number of defaults, total number of accounts, maximum and minimum of certain variables are derived. The total values will be required to calculate the WoEs in the grouping stage. The bins are assigned at the mapper stage and the derived bins are the keys which feed into the reducer stage. In these stages the total number of defaults is counted. The same reducer is used repeatedly for each grouped WoE calculation. This reducer has two mapper inputs. The

total input and the output from the respective grouping mapper.

Given that a small number of bins are present, it is not considered inefficient to store the derived WoEs in memory. The calculated WoEs were then applied to the input data and the ABT for the Logistic regression trainer was created.

The resulting dataset was then trained using a logistic regression also built using Hadoop map reduce. The logistic regression code used for this was leveraged heavily off the logistic regression mapper and reducer. The code uses the gradient descent algorithm to optimize the coefficient values. The code such that the training set is taken from the HDFS, and a logistic regression is trained given a learning rate parameter α and the number of times to iterate over the training set as inputs.

RESULTS

After the analysis, the following results and conclusions were obtained and they are listed below:

Feature	Bin	Default rate	Weight of Evidence (WoE)
LTV and Portfolio	Mortgage: LTV less than 113%	6.56%	0.8362
	Mortgage: LTV between 113% and 120%	8.64%	0.5394
	Mortgage: LTV greater than 120%	11.71%	0.2005
	Vehicle Loan: LTV less than 66%	15.01%	0.0854
	Vehicle Loan: LTV between 66% and 78%	20.47%	-0.462
	Vehicle Loan: LTV greater than 78%	25.65%	-0.755
Monthly Instalment	Greater than 750	8.33%	0.5797
	Less than 750	21.02%	-0.4953
Age	Less than 35	18.22%	-0.3175
	Between 35 and 55	12.43%	0.1328
	Greater than 55	6.50%	0.8419
Normalized External Score	Less than 33% (Bad or no score)	23.10%	-0.6164
	Greater than 33% (Good score)	10.22%	0.354

Table 1 Experimental results

Conclusion 1.: Vehicle loans have a much higher default rate that Mortgages. This was to be expected however the difference in default rates was surprising. For Vehicle loans a default rate of ~20% was observed while for mortgages the default rate was around 8%. This information would be very useful for banks as it would mean that it would be beneficial to apply varying credit policies between mortgages and vehicle loans.

Conclusion 2.: The older the applicant is the less likely they are to go into default. This is an interesting insight and using this information, banks can use this to target older customers. Banks can also use this to charge a higher default premium on credit given to younger customers.

Conclusion 3.: The loan to value (LTV) distribution between mortgages and vehicle loans was so different that the model had to be adjusted to accommodate this. The average LTV for vehicle loans was much lower than that for mortgages. This could be attributed by the fact that Vehicles tend to depreciate a lot faster. This gave problems for the model as LTV is associated with having a positive relationship with default risk. Given that Vehicle loan default rates are much higher than Mortgage default rates, binning by LTV alone would not result in monotonic behaviour. To remedy this, the decision was taken to group the LTV and Portfolio variables together.

Conclusion 4.: External credit scores do play a role in measuring credit quality. The best way to bin external credit score was to split between normalized credit scores of above 33% and below 33%. A more granular split between external scores was expected

from this analysis however it could be argued that this still is useful in determining credit quality.

The resulting regression all had coefficients with negative signs. This was as expected given that WoE has a negative relationship with credit defaults. The chosen learning rate parameter was 0.05 and the training set was iterated over 400 times.

CONCLUSION

Banks and Financial Institutions are heavily exposed to default risk. Default risk is the inability of obligors to meet their contractual repayments on their debt obligation. Default risk is the largest operational expense that banks face. Banks become insolvent when many of their obligors cannot meet their contractual repayments. Because of the adverse economic effects that can result from banks facing default risk, there exists a lot of regulatory pressure on banks to build effective and explainable models. Some of the main conclusions of this research paper are:

- Vehicle loans have a much higher default rate that Mortgages;
- The older the applicant is the less likely they are to go into default;
- The loan to value (LTV) distribution between mortgages and vehicle loans was so different that the model had to be adjusted to accommodate this;
- External credit scores do play a role in measuring credit quality.

This research paper might become useful if banks ever do decide to use big-data technologies. Some valuable insights were derived from this analysis and they could be used to drive important banking business decisions.

REFERENCES

1. A. a. J. Bravo, "A Non-Parametric-Based Computationally Efficient Approach for Credit Scoring," 2019.
2. V. T. Pavel Mironchyk, "Monotone optimal binningalgorithm for credit risk modeling," 2017.
3. Abdou, Hussein A., and John Pointon. "Credit scoring, statistical techniques and evalua-tion criteria: a review of the literature." Intelligent Systems in Accounting, Finance and Management 18.2-3 (2011): 59-88.
4. Lessmann, Stefan, et al. "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research." European Journal of Operational Research 247.1 (2015): 124-136.
5. Onay, Ceylan, and Elif Öztürk. "A review of credit scoring research in the age of Big Da-ta." Journal of Financial Regulation and Compliance 26.3 (2018): 382-405.
6. Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." Communications of the ACM 51.1 (2008): 107-113.
7. Pop, Daniel, Gabriel Iuhasz, and Dana Petcu. "Distributed platforms and cloud services: enabling machine learning for big data." Data Science and Big Data Computing. Springer, Cham, 2016. 139-159.
8. Liu, Zhen, and Meng Liu. "Logistic regression parameter estimation based on parallel ma-trix computation." International Conference on Theoretical and Mathematical Foundations of Computer Science. Springer, Berlin, Heidelberg, 2011.
9. Chu, Cheng-Tao, et al. "Map-reduce for machine learning on multicore." Advances in neu-ral information processing systems. 2007.
10. Zaharia, Matei, et al. "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing." Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation. USENIX Association, 2012.
11. Chowdhury, Mosharaf, Matei Zaharia, and Ion Stoica. "Performance and scalability of broadcast in Spark." press. 2014.
12. Zaharia, Matei, et al. "Spark: Cluster computing with working sets." HotCloud 10.10-10 (2010): 95.
13. Rousseeuw, Peter J., and Katrien Van Driessen. "A fast algorithm for the minimum covar-iance determinant estimator." Technometrics 41.3 (1999): 212-223