HOME CREDIT DEFAULT RISK DETECTION

This paper discusses the business use case and need for a machine learning algorithm to assess the Home credit default risk of a borrower.

MAP REDUCERS

Course: BUDT758B, Spring 2020

Big Data and Artificial Intelligence for Business

1. Understanding Home Credit Default Risk

Home Credit Default Risk has been a prevalent concern in the housing sector for decades. Credit risk is defined as "the possibility of a loss resulting from a borrower's failure to repay a loan or meet contractual obligations."[1] This risk is especially important in the housing and financial sector where a default can result in a financial institution not being paid back. That can result in the financial institution not being able to meet their own obligations or carry out their own operations. In addition to financial loss, defaults can also result in the financial institution not extending credit to other borrowers, which can have a broader impact on the economy.
While banks extend credit to many borrowers, there are other organizations that can lend money to borrowers. The most common places that extend credit are banks, credit unions, payday lenders, pawn shops and cash advancement companies. Credit can also be found more personally from family, friends and one's own 401k[2]. Since so many different people and institutions can lend money, credit default risk is more widespread than just the banks that most people may be accustomed to.

Credit risk is widespread and present with every borrower to varying degrees. Some borrowers are certainly more credit worthy than others, but that does not mean there is such a thing as a risk-free borrower. No matter how safe a borrower is, there is always risk. Credit risk is present with every borrower and can result in devastating losses to an organization. For example, according to CNBC, more than 1 million borrowers' default on their student loans each year[3]. According to the same article, almost 40% of borrowers are expected to default on student loans within the next 3 years. Many of these loans are federal and offered by the federal government. The remainder of the loans are private and offered by various lending institutions.

---

[1] https://www.investopedia.com/terms/c/creditrisk.asp

[2] https://www.creditkarma.com/personal-loans/i/where-to-borrow-money/

[3] https://www.cnbc.com/2018/08/13/twenty-two-percent-of-student-loan-borrowers-fall-into-default.html

2. Business Impact of Home Credit Default Risk

The effects of these defaults can be felt by nearly every stakeholder of the organization. The borrower can have their credit ruined by the default[4]. This means they will be seen as less creditworthy and may not be able to receive loans in the future. This can make it difficult for them to buy a car or a house in the future. This default can also result in garnishment of the borrower's wages[5]. This means that the government can give the lender the authority to take wages out of the borrower's paycheck without the borrower's consent. This means the borrower will have less money available to meet obligations, pay bills and handle any emergency costs. The effects of a default are also felt by the lending organization itself. Since the borrower cannot pay back the loan, the lender may not be able to recover all of the money. The lender can possibly perform a wage garnishment, but that does not guarantee the lender will receive all of their money. In the case of Home credit default, it leads to foreclosure of property by the bank. The lending institution may have to sell the property at a lower than market price. This can hurt the lending institution's bottom line and result in the loss of faith in the company by its shareholders.

The shareholders of the lending institution are also hurt by the inability of the lending institution to properly detect and handle credit risk. The value of their assets can decrease causing them to lose wealth. All of the stakeholders of the company are hurt. Even the employees are affected because the company's ability to pay their employees could be affected. A large amount of defaults could result in insolvency on behalf of the lending institution. This could result in the company having to perform mass layoffs or even closing their doors for good. This is what happened in the 2008 financial crisis when certain financial institutions could not meet their financial obligations[6].

---

[4] https://loans.usnews.com/what-happens-if-you-default-on-a-loan

[5] https://loans.usnews.com/what-happens-if-you-default-on-a-loan

[6] https://www.investopedia.com/articles/economics/09/lehman-brothers-collapse.asp

3. Need for a Solution & Our Approach

For a very long time, banks and other investment institutions have hired many analysts in order to detect credit risk. With the big data revolution and the advent of data science, a more quantitative approach can be taken to the issue. That is what inspired our team to tackle this issue. The issue of Home credit default risk has a historic and well-documented business case with implications involving the general public, lenders, borrowers, employees as well as government institutions who may be responsible for backing deposits that are loaned out.

3.1 Extract Transform Load (ETL)

Our team, **Map Reducers**, has acquired data about various credit cases from Kaggle, an online data science competition facilitator and data repository. The first thing our team did was build an ETL pipeline before we could analyze our data. We began our data extraction by extracting our data from Kaggle. We used Google Colab as our development environment. We were able to load our data directly from Kaggle into Colab and unzip the files. Once we had access to our data in the form of a dataframe, we had to transform it into something usable. We had to merge several different data files in order to create one file to be used in our analysis. We also had to remove several outliers in various columns that did not make sense given our domain knowledge. The final step in our ETL process was to save our prepped data, to avoid having to prepare data every time we opened the notebook. We decided to save our data to Google Drive because it is free and promotes collaboration. This concluded our ETL process.

3.2 Building Predictive Models

Once our data was properly merged and cleaned, we were able to begin our analysis. We ran several different models and compared the results in order to choose the best one. Since the data had 92% of non-defaulter's applications and just 8% of defaulter's applications, the baseline accuracy of our training set was initially 92%. After running logistic regression which predicts well with a linear boundary and random forest which uses a non-linear boundary, our accuracy

never went past 63% and true positive rate was below 20%. Hence, we had to under sample the non-defaulter's application records such that our training set had equal number of defaulters and non-defaulters (50:50) application records and our test dataset had the extreme bias just like in the real-world datasets.

We then ran k-means clustering, logistic regression, deep neural network, random forest and XGBoost model on our training set. Each model gave us a glimpse into the nature of our data. Initially, our assumption was that there were identifiable clusters within the data. Hence, we ran the K-Means algorithm on our data. However, the clusters obtained were not very homogenous, which led us to believe that the data was not easily separable between the two classes. This led us to run a logistic regression model. We then discovered that the data was not linearly separable even with the right class weights, and thus had to use algorithms which have a non-linear separable boundary. We confirmed this by running Principal Component Analysis on the training dataset. This led us to run a random forest model. Also, upon discovering how complicated this dataset was, we decided to try a more complex approach. We ran a deep neural network with convolutional and dense layers to improve our predictions. We understood that the deep neural network was not able to learn much although it gave good enough metrics. Finally, we decided to build upon our random forest model by using XGBoost, which improved the model's accuracy and gave us a decent true positive rate of 70.11% along with a good accuracy of 70.47%.

Surprisingly, these figures were very similar to that of the logistic regression model. We decided that we would use the XGBoost model for future predictions because it's less likely to overfit and does automatic feature selection. Correctly identifying the defaulters is the priority here, so it makes sense to use the true positive rate as an important metric along with accuracy.

3.3 Key Findings & Insights

Our analysis of this data has led us to several findings. The dataset we used was highly biased. 92% of the data was of non-defaulter's applications. Defaulted transactions are rare given a large pool of transactions. Something interesting that we found in the data was that the source of the transaction had a lot of predictive power to determine whether or not the borrower was likely to

default or not. This information can be used to predict defaults and therefore reduce the amount of defaults in an organization.

Reducing the amount of defaults in an organization can benefit the employees, customers, constituents and every other stakeholder. Our results would be interesting to company executives who are interested in reducing the amount of defaults within their organization and to those who work within the government. It is essential that the government be able to maintain financial solvency. Preventing defaults is an important step in doing that. That holds especially true for the housing and financial services industry.

4. Conclusion & Next Steps

In conclusion, default on home credit loans is a devastating problem that can wreak havoc on an organization's customers, employees, profits, shareholders and stakeholders, as well as the general public. Our team has created an advanced analytics solution to assist organizations in detecting Home credit default risk so that they can respond appropriately to it. This solution was created in Python with the goal of using machine learning to detect potential defaulters. This can result in organizations having more engaged employees, more financially healthy customers and higher profits. This will improve the lives of lenders, borrowers, shareholders and stakeholders.

As a scope for further research we believe in continuous improvement of our model to meet varying business requirements. For instance, the current COVID-19 pandemic may well lead to home owners defaulting on their mortgages. They may be furloughed or simply be unable to reach their workplace for work. This could also lead to a slide in home prices. So, an Ideal Model should account and adapt to such extreme situations in a short span of time. Given the limited data availability at this time, it could teach itself to learn from this sudden change in borrower pattern and make post pandemic predictions in the housing market.