1. Problem Statement

Over the last 20 years, bank failure has plagued the financial system more so in emerging markets (Bell and Pain, 2000). Llellewyn (2002), states in his paper that an inefficient process for risk analysis, management and control is one of the most significant factors leading to bank failures.

Banks face 8 main risks: credit risk, market risk, operational risks, moral hazard, liquidity risk, business risk, reputational risk and systemic risk. Of these, credit risk is one of the greatest risks inherent in the banking business. Credit risk refers to the uncertainty that banks face which arises from the possibility of non-payment by borrowers. It can adversely affect a bank's profitability and cash flows thus it needs to be closely monitored and managed.

This research attempts to build a model that estimates cash flows to a bank whilst considering the credit risk of individual borrowers using a credit scoring model. Such a model would be beneficial to banks as it enables them to predict their cash flows and consequently run their businesses efficiently.

Current Issues

After the implementation of the interest rate collar, banks shifted away from lending to customers and instead invested in government paper. The oversubscription on Treasury Bills and Bonds has effectuated declining yields on government paper thus pushing banks to bulk up their loan books in order to remain profitable. The latter however increases their exposure to credit risk. As a result, banks are forced to be more cautious when it comes to managing credit risk and considering loan applicants in an environment where default premiums are constrained.

2. Data Description

i.) General Observations

a.) Subject of the dataset

The data set details 500,000 loans that have been extended to 266,983 bank customers. The data contains 14 features which describe the customer and the borrowed amount.

b.) Null Values

The data set contained some null values (11 observations across 10 variables). These were handled in the data cleaning section.

c.) Data Types

The Age, CustomerType, SOR, MonthsSinceOpen, MonthsSinceActive and the CRBScore variables were initialized as float variables as opposed to integers however this isn't much of an issue since the data is numerical.

d.) Column Names

Columns were consistently named except the Customer ID column which was changed.

ii.) Desciptive Analysis

Identifier Variable

1. CustomerID

This variable is a unique customer identifier. As mentioned above, the data set includes 266,983 customers.

Numerical Variables

1. DisbursementDate

This variable details the date when the bank disbursed the loan. The dataset contains loans disbursed between 3rd January 2017 and 30th November 2017.

2. Age

The average age of customers is 38 years at the time of taking up a loan. About 68% of customers are between the ages of 27 and 49. The youngest customer is 3 years old whilst the oldest customer is 118 years old. The average age is slightly above the median of 36 indicating that the data is slightly positively skewed which could be the effect of positive outliers (the 1,239 customers aged 117). The minimum and maximum values are very far from the lower and upper quartiles thus suggesting that this variable contains outliers. The banks loan book consists of a relatively young borrowers.

3. MonthsSinceOpen

This variable describes the number of months since the customer first opened an account at the bank to the date the loan was disbursed. During the period, customers that received loans from the bank had on average been customers of the bank for approximately 90 months (7.5 years). However, the bank also issued loans to customers that had only been with the bank for only 4 months. The customer with the longest history with the bank who received a loan during the period had been with the bank for 144 months (12 years).

4. MonthsSinceActive

The bank issued loans to relatively active customers. On average, customers who received loans were those whose last activity was less than a month before the loan was disbursed. An exception was however made with 1 customer who was issued a loan despite the customer's last activity being 2 month prior. This variable was highly skewed ($\gamma 1 = 707.1$), indicating that the bank highly favors customers who have banked with them for a while.

5. FinancialMeasure1

This measure describes some score of customer activity. It ranges from 0 to 2,186,100 although 67% of the observations are zero. Borrowers on average attain 15,360 of this metric. The 95-th percentile is 71,616 whereas

the maximum value is 2,186,100. The large dispersion between this and the maximum value is indicative of outliers.

6. FinancialMeasure2

This measure describes some score of customer activity. It ranges from -4,523,500 to 5,705,200. Close to 1% of the data is zero. On average, customers that receive loans attain 11,905 of this measure. The 95-th percentile is 45,238 and the large dispersion between this and the maximum value (5,705,200) is also indicative of outliers.

7. FinancialMeasure3

This measure describes some score of customer activity. It ranges from 0 o 97,479,000. At the time of application, customers had a financial measure 3 value of 100,920. Further, the measure is heavily skewed ($\gamma 1 = 42.571$). The 95-th percentile is 265,580 far lower than the maximum value. The latter indicates that the data set has extreme outliers

8. FinancialMeasure4

This variable was dropped from the analysis since it was highly correlated to financial measure 3.

9. CRBScore

On average, customers had a credit score of 586 at the time of loan application. This however varied by 110 across customers. A loan was however issued to a customer with a credit score of 0. Most customers had a credit score less than 653 at the time of application. The customer with the highest credit score who got a loan had a score of 747. The wide range between the minimum value of 0 and the lower quartile- 547 is indicative of outliers.

10. Amount

The average loan amount issued is 9,155.10. This however varies by 16,812.81. The lowest loan amount issued is 100 whereas the largest loan disbursed is worth 200,000. However, most of the loans are worth less than 9,900. The latter is also indicative of outliers as the gap between the upper quartile and the maximum value is very large.

Categorical Variables

1. CustomerType

72.1% of the loans issued were issued to customer type 1 customers whereas 27.9% of the loans were issued to customer type 2 customers.

2. SOR

This is a metric of customer relationship with the bank. A majority of the loans were issued to customers with an SOR of 5 (62%). Customers with an SOR of 5,4,3,6 and 7 received the majority of the loans with the rest of the groups receiving less than 1% of the issued loans.

3. Default

The default rate on loans is very low with only 2.4% of the loans registered as defaulted. The latter means that the variable is heavily imbalanced and would need to be balanced before proceeding wih the analysis.

3. Data Exploration

a.) Correlation Analysis

None of the variables is heavily correlated with another except Financial measure 4 and 3 which was mentioned earlier.

Default is slightly negatively correlated with age, SOR, months since opened and the CBR score. The latter means that as the latter variables decrease the more likely it is that the customer has defaulted on the loan.

The SOR and the amount of the loan are also positively correlated. The latter would b=make sense if a greater SOR score means that the bank has a better relationship with the bank.

The financial measures, customer type and CRB score are positively correlated with the amount lent. The latter means that an increased financial measure, customer type and CRB score increases ones chances of receiving a greater loan.

b.) Customers

i.) Customer Type

Customer Type has the highest relation with Financial Measure 1. This indicates that customers across the different types share have relatively similar features used to calculate Financial Measure 1. This is likely because the customers are grouped based on the underlying metrics of this measure. For example high net worth individual and customers with lower net worth. There are 4 types of customers however, customer type 1 and 2 take the most loans forming 72.07% and 27.93% of the loan book. Further, the cross tabulation also shows that 97.61% of the loans taken by customers have not been defaulted where as 2.38% of the loans have been defaulted. From the visualizations it appears that there is a difference in default across different customer types. Customer type 2 has the highest default rate a 3.7%. This customer group should be investigated to reveal what makes them more prone to default. The bank could take feedback from research and created tailored products for this group so as to manage default and serve the customers better.

Hypothesis:

There is a statistically significant difference in default across different customer types.

ii.) SOR

There are 8 rankings describing the customer relationship with the bank. The SOR has the highest correlation with the amount disbursed. This makes sense as it is easier for a bank to issue higher value loans to a customer with whom the bank has a good relationship with. Customers with an SOR of 5 dominate the loan book claiming 62.2% of issued loans. On the other hand, customers with an SOR have the highest default rate (0.69%). From the visualization it appears that there is a difference in the default rate across categories.

Hypothesis:

There is a statistically significant difference in the default rate across categories.

c.) Financial Measures

On average, the financial measures are higher for loans that were paid back as opposed to defaulted loans. The latter indicates that the financial measures are positive indications of credit worthiness for example net worth. From the visualization it appears that there is a difference in the financial measures across the default variable.

Hypothesis:

The hypotheses for the 3 variables is therefore that there is a statistical difference in the average financial measures across the default variable.

d.) Amount

On average, there doesn't seem to be much of a difference between the amount borrowed for loans that were paid back and those that were defaulted. The average amount of loans defaulted however is slightly higher.

Hypothesis:

There is no statistically significant difference in the loan amount across the default variable.

e.) Months Since Open

Inferring from the visualization, loans that have been defaulted are on average requested by customers who have been with the bank for a shorter period of time.

Hypothesis:

There is a statistically significant difference in the duration of membership across the default variable

f.) Age

From the visualization, we can tell that customers who default on loans are on average slightly younger than those that don't default.

Hypothesis:

There is a statistically significant difference in the age of borrowers across the default variable

g.) Trends

i.) Daily Trends

I investigated the average daily amount disbursed, the average number of loans disbursed, the average daily default rate and the average daily age of borrowers. The average daily loan amount disbursed peaked on the 14th of June. The Default rate on loans issued on that day were however mostly paid back- only 1.7% were defaulted. Of the defaulted loans on this date, borrowers outperformed the average defaulter statistics across the financial measures except financial measure 3 where the average came in at 52,841 as opposed to the defaulter's average of 60,000. Loans on this specific day may need further analysis to rule out suspicious activity. The average daily default rate eased towards the end of the period as the average age of borrowers increased slightly.

ii.) Monthly Trends

The daily trends are very noisy and relatively difficult to pick out trend however, thus a monthly trend analysis was also carried out. The monthly average loan amount disbursed shows peaks in January and May and drastically in April and September. It would be interesting to see trends of this variable over many years so as to establish if the trends are cyclical.

The Average number of loans peaked in March and has declined since then. The average age was also at its lowest between February and April while default rate was at its highest in March. It would be interesting to evaluate what events happen around the same time and are peculiar to the youth in order to determine what could drive the increased loan during the period.

4. Data Cleaning

a.) Dropped Variables

As mentioned, financial measure 4 was dropped owing to its high correlation with Financial Measure 3

MonthsSinceActive was also dropped since approximately 100% of the series values were 0. The variable therefore provides little information in the analysis.

b.) Duplicate Values

Analysis of the data set found that there were 524 duplicate values which were dropped to avoid bias in the model.

c.) Missing Data

The 'Age', 'CustomerType', 'SOR', 'MonthsSinceOpen', 'MonthsSinceActive', 'FinancialMeasure1', 'FinancialMeasure2', 'FinancialMeasure3' variables all had 11 missing values. Upon further investigation, it was evident that it was the same observations that contained the missing variables.

5 of the 11 customers had previous data that would have been used to replace current values using the code defined below. However, this method assumes that a customer's financial metrics CRB score, customer type and SOR are constant over time. The latter would have introduced some level of bias to the model. I therefore opted to drop the null values since they were not significant.

d.) Dealing with Outliers

This study used a combination of techniques to treat the outliers. For most of the columns, I used the z-score method to determine outliers- values that were more than 3 standard deviations away from the average- above the 99.7th percentile and below the 0.3 percentile.

To treat outliers, I first set the outliers to null then I imputed values for the latterly defined outliers based on other correlated fields. For example, Financial measure 1 had the highest correlation with SOR. The Financial Measure 1 values were first set to null the grouped based on the SOR. The null values were the replace by the average value for Financial Measure 1 in their respective SOR bins.

Financial Measures 2 and 3 had the highest correlation with the loan amount thus a categorical variable- Loan Category was created based on Loan Amount. The Financial Measure 2 and 3 outliers were then set to null and grouped by the loan category. Finally, the null values (outliers) were replaced by the null values of their respective bins.

Outliers for the 'Age' variable were determined by visual analysis and logic. It is illegal for a person below the age of 18 to enter into a contract thus it is not feasible that they received a loan.

The treatment of outliers was also not very harsh as tail risk can greatly affect a banks income. Further, to create a model robust to outliers, I chose to use tree based methods that aren't as affected by outliers as regression based models e.g. the logistic regression.

e.) Appropriate Data types

I changed the 'CustomerType', 'SOR', 'LoanCategory', 'CustomerLevel', 'DisbursementDate_month', 'MonthTime' variables to the categorical data type

f.) Balancing the Data

The data is heavily imbalanced thus some action should be taken towards rectifying te latter. An under sampling technique would have truncated a lot of the data thus an over sampling technique was first run. The latter however worsened the model thus a stratified sample was used. Machine learning algorithms often assume that the distribution for the training data is the sample in the test set thus stratifying the samples aids in improving prediction accuracy.

This steps were done just before modeling.

5. Feature Engineering

a.) Disbursement Date

I chose to split the disbursement date into monthly and daily data so I could better investigate trends in the disbursement date of loans. For example are people who were issued a loan in January more likely to default as opposed to people who were issued the loan in May? Similarly, I also wanted to see if the timing of the disbursement within the month affected default. I further categorized the days into either the beginning of the month (1), Mid-Month (2) and the end of the month (3).

b.) Default

I introduced a feature called Default Rate which summarizes a customer's previous default history. The metric simply takes an average of previous loans taken by a customer at the time of the loan disbursement. It therefore changes over time and gives a better picture of a customer's default patterns as we get more default information on the customer. I used the expanding().mean() functions to find an incremental average of default with respect to time for each customer. The latter combination enabled me to calculate the average default for each customer based on the pervious and the current loans. However, the variable wasn't realistic since it included the current loan default- a metric which the bank doesn't have at the time of disbursement for the current loan. I therefore used the shift() function to move the variable forward by one time period for all customers so that the default rate only included previous default information at the time of disbursement. The variable had to be encoded to allow for analysis on the variable.

c.) Average Financial Measures to Date

As with the default rate variable, I also calculated the increasing averages for the financial measures columns. The columns basically summarize the average of the financial measures at the time of disbursement of the current and previous loans since the bank has access to this measures on the date of disbursement.

d.) Average Amount and Amount Difference

Similarly, a column was created which summarizes the average loan amount the customers borrows up to and including the current loan. As with the other metrics the values changes for each customer as they borrow more and more.

In addition, a variable call AveDiff was created which summarizes the dispersion of the current amount borrowed from previous amount borrowed. This feature raises flags for amounts that are significantly higher than a customer's previous loans and allows the bank to monitor such loans.

6. Feature Selection

SelectKBest

The Financial Measure 2 metrics and the average amount difference were disregarded from the initial chi2 test since the test assumes that the variables are distributions and therefore non-negative. The features were later re-introduced into the model and the model was evaluated to see if the improved or worsened the model. I used the chi-square test of significance since it tests for independence/dependence between the y variable and the X variables.

The model with Average Amount Difference in addition to the chi-square parameters yielded the highest precision results for the defaulters thus the variables were used in estimation