Oliver Stewart

Capstone project

Bank Default Rates

Contents

[Relevance 2](#_Toc104030066)

[EDA, Feature Selection and Understanding the Data 2](#_Toc104030067)

[EDA 2](#_Toc104030068)

[Feature Selection 3](#_Toc104030069)

[Understanding the Data 3](#_Toc104030070)

[Modelling 4](#_Toc104030071)

[Expectations 4](#_Toc104030072)

[Assessment 4](#_Toc104030073)

[Concluding Remarks 5](#_Toc104030074)

[Column Input explanation 7](#_Toc104030075)

[Feature Importance 8](#_Toc104030076)

# Relevance

This capstone work is based on a Kaggle sourced from the U.S. Small Business Administration (SBA). The SBA is an organisation that aids small organisations in the credit market. The dataset lists over 800,000 loans that small businesses have undertaken ranging from 1968 to 2014. The loans have columns including the location of the loan, the industry the loanee belongs to, how much is being loaned.

When framing what could be done with this data and how it could create value for a stakeholder, we see a variable that is based on if the loan defaults or not. This is ideal for banks who would like to know the characteristics of those loans that default and if it is worthwhile to grant a loan to a business. This creates the question of **“Can we create a model that predicts if a loan will default or not?”** This can bring up follow up questions including “What are the most important factors in determining a defaulted loan?”

Banks are in the business of creating a return on their investment and a large majority of businesses are classified as small, meaning these loanees are not niche and constitute the common entities that require a loan. This reinforces the importance to be proficient in the context of dealing with these kinds of businesses.

# EDA, Feature Selection and Understanding the Data

## EDA

In its initial stage, the data is not usable as many of the columns are unusable as they are not in numerical form, so this needed to be converted. This could be done once unneeded columns were dropped as the question framed had no use for what state the bank was located for example (the loan is defaulted based on the loanee not the loaner).

Columns were converted to their desirable type (converted from an object to a float) and some columns were dropped due to being very similar to other columns such as the state, city and zip codes of the loanee. In this instance the state feature was kept as it would provide the least number of columns when splitting it into dummy variables and still gave a locational input. Inputs additionally had unneeded characters in them such as the monetary figures where ‘$’ and ‘,’ and extra characters on dates which needed removing.

Columns that also were slight alterations of each other included the amount disbursed and the gross amount on a loan. Most of these values were the same as the amount promised is typically the amount given unless waiting on extra instalments, so only the gross amount was kept.

Observations with null values were removed from the dataset and dummy variables were now put in with the largest inclusion of dummy variables came from the State and Industry of each row. The industries aligned with the North American Industry Classification System and gave distinct industries each loanee worked in that could be identified with the first two digits in the NAICS code. This led to mapping the industries to these digits and then creating dummy variables for them.

The last step to clean the data included mapping values so they were 1s and 0s and dropping those last columns with values that could not be converted numerically.

Throughout creating these dummy variables, we need to have bases as including all variables in a dummy split would lead to multicollinearity issues which is fixed with dropping one of the dummy columns. This becomes the reference category, for this project the reference category was the most common dummy which included:

* State: California (CA)
* Industry: Unspecified
* Area Classification (Urban/Rural/Undefined): Undefined

## Feature Selection

A screenshot of a computer

Description automatically generated with medium confidenceWhen creating a correlation matrix for the inputs, there was a low magnitude of correlation for most values including the target variable, meaning that there was likely a complex number of inputs which led to a loan defaulting or not. With no apparent relations it was more of a task to understand the distribution and makeup of the data. With a lack of any input with collinearity issues, the remaining features were kept which amounted to 81 columns (80 inputs and 1 target variable).

## Understanding the Data

In terms of the actual distributions some created histograms that had binning issues due to a wide range of data and many outliers being present which didn’t give the best representation of the data. For instance, the range of number of employees were from 1-9999 but the majority had businesses with less than 5 employees. To get a greater idea of the true distribution we created a filtered histogram that included cases with less than 100 employees highlighting the frequency in the low numbers.

|  |  |  |
| --- | --- | --- |
| **Column** | **Description** | **Graph** |
| Year | The data is more skewed towards recent years, could hint to a habit of loans being more common now as well as the prevalence of more businesses/people living in the U.S. | Chart, histogram  Description automatically generated |
| Term Length | Terms tend to be 100 months or less which is more indicative of a smaller business | Chart, histogram  Description automatically generated |
| Amount of Loan | The loan amount these businesses take are usually at most in the tens of thousands range which is indicative of the size of a business | Chart, histogram  Description automatically generated |
| Employee  Number | This is filtered to only show observations with less than 100 employees but can see the large majority have less than 10 employees (500K+/800K rows) | Chart, histogram  Description automatically generated |

Other relations and potential trends were attempted to be discerned like loan length vs loan amount, but no clear conclusions could be drawn. However, the identity of the data was clearer as the distributions of the data indicate that the observations line up with characteristics of small businesses. As there were no strong relations with the target in the correlation matrix, modelling techniques will need to be implemented to generate better predictive power.

# Modelling

## Expectations

The next stage is onto modelling the data. With the evaluation of the model, we need comparison. The baseline accuracy of the data is 82% which means if we guess every loan will be paid, we will be right 82% of the time. The key is not the accuracy figure of a particular model but the precision and recall of how well it identifies a defaulted loan.

Recall is the model’s ability to find all the relevant cases within a data set whilst precision is the ability of a model to identify only the relevant data points. This keeps in mind the original question we have formulated stating emphasis on detecting these defaulted loans as the banks want to be risk averse meaning it is worse to accept a loan that would default than to reject one that won’t in the hopes of lowering credit risk.

The process of the models will include using logistic regression, decision trees, random forests and gradient boosting (including XGBoost). If the model improves, ensemble techniques will be applied. The models will then

## Assessment

By utilising all the modelling techniques and considering any applicable ensemble modelling we ended up with these figures:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Default Precision | Default Recall |
| Logistic Regression | .85 | .69 | .25 |
| Decision Tree | .94 | .85 | .81 |
| Random Forest (Bagged) | .95 | .87 | .75 |
| Gradient Boost | .93 | .85 | .73 |
| XG Boost | .95 | .87 | .82 |

From these models we can see that the accuracy improved on the baseline of 82%. It is impressive to see a rise in the default detection with the most successful techniques get a recall of up to 82% which in the context of this scenario is the most important measure as Banks are more concerned with those that would default and go undetected as that loss is the most impactful.

In terms of the ability to identify the loans that would be repaid all models did exceptional with all but 1 model (logistic regression) scoring in the mid-high 90s in both precision and recall. The models do better at identifying the paid loans than the defaulted ones mostly due to the underlying fact that the data includes these paid loan observations at a more frequent rate.

The ability to still correctly classify these paid loans means the model still works the other way where business and loans can be made as there is an inevitable trade-off. You can accept more loans through which opens more credit risk. Conversely, you can deny more loans to mitigate risk but will fail to generate revenue. These models aim to reduce this potential loss that comes with the trade-off both ways and from these figures of the better performing models, that is achieved.

A picture containing line chart

Description automatically generatedThe other way we can evaluate these models is the ROC area under the curve and the results are also impressive with high scores as well.

Upon assessing the models, it is determined that the final one to be used is the XGBoost model. With its strong predictive capabilities, it answers the question of whether we can predict if a loan will default or not to a high degree. The follow up question is what the more impactful inputs for are for whether a loan defaults and the main contributor is the term length of a loan. This is the biggest factor by a considerable margin next being the year the loan was approved. A takeaway from the feature importance is that the state the loanee is in typically has a larger impact than the industry the business is from.

**Note: The explanation of features will be included in the end of this documents as well as the feature importance graph.**

# Concluding Remarks

From the modelling conducted, a strong model was created that can effectively determine if a loan will default. This provides tangible value for the banks who can determine if it is worth granting a loan to a business given certain criteria.

As seen with the amounts of the loans generally distributed which were in the tens of thousands range. Being aware of the state of a business you are dealing with aids loaners to be informed on decisions before providing capital. As small businesses can be unknown with little proprietary information, banks can struggle to ascertain whether the business is sustainable and will repay what is owed and through this model that problem is improved.

Regarding the use of these models in a practical environment, they should be a supplementary tool in measuring the ability of a business to repay its debt. The model should not be the only screening tool used as it is not perfect but still performs well. It is based on the idea that information for smaller scale businesses is harder to acquire but lenders must do due diligence and enquire beyond the needed inputs.

The model can also be used for not just loaners but also those who guarantee loans. As the basis of the model is on the characteristics of the loanee themselves, external parties have no direct influence on the prediction of the model. Guarantors are relevant stakeholders in loan transactions and would be available to use this model to aid in the decision to guarantee a loan or not.

# Column Input explanation

|  |  |
| --- | --- |
| Input Total= | Description |
| MIS\_Status (Target) | 1 = Defaulted Loan 2 = Paid Loan |
| Term | Length of loan (months) |
| Approval FY | Year loan was approved |
| Urban | In Urban area |
| Rural | In Rural area |
| GrAppv | Gross amount of loan approved by Bank |
| NewExist | New Business = 1 Existing Business = 0 |
| RetainedJob | Number of Jobs Retained |
| CreateJob | Number of Jobs Created |
| Franchise | Franchise = 1 Not a Franchise = 2 |
| NoEmp | Number of Employees |
| Two Letter word i.e. AZ | A certain US State i.e. Arizona = AZ |
| Else | NAICS labelled industry |

# Chart, histogram Description automatically generatedFeature Importance