Predicting Loan Eligibility of SMEs in California

Logistic Regression and Random Forests

Abstract:

Qualifying for loans is essential to all businesses - especially Small and Medium-sized Enterprises (SMEs). The birth and continuation of many SMEs depend on their ability to acquire a loan. Financial institutions, on the other hand, are tasked with estimating an amount they believe the SME can return so as to minimize the default rate. This article reports the results of supervised and semi-supervised modeling techniques used to simulate the processes financial institutions use to determine loan amounts and eligibility. These techniques include Logistic regression models and a tree classification model. The dataset is derived from the Small Business Administration (SBA), which includes historical data from SMEs in California spanning nearly 30 years We run our models on the data, report our results, and determine which type of model works best for this use case. We hope that our findings can be extrapolated to SMEs all over the country.

1. Background and Significance:

Small and Medium-sized Enterprises (SMEs) account for 99.9% of U.S. employer firms and are an integral part of the economy [1]. As of October 2020, SMEs consist of 31.7 million businesses in the U.S. and generate two-thirds of all new jobs in the country [2]. We wanted to understand the process behind the SBAs approving loans for small businesses because they are integral to our economy and society. The mom-and-pop stores, immigrant-run ethnic grocery stores, and home bakery businesses in your local neighborhood are SMEs. Their contributions to and participation in communities are among the many reasons why we should protect them from shutting down.

SMEs depend on loans from financial institutions to start and support operations; loans prove especially crucial in times of credit business transactions, which many do. The SBA is a government agency that aims to support SMEs around the country. Understanding how the SBA approves loans and using predictive models can help SMEs estimate the loan amount they would be approved. More accurate loan expectations can then translate to smoother financial planning and help the SBA minimize its risks.

Based on the literature review, we hypothesize that Random Forest models will perform well with this data because of their sequential classification tree method [3]. Additionally, the use of Logistic Regression models seems especially popular within the loan default prediction literature. With this type of regression, we can discover which properties were most directly related to the chance of loan default and which properties could be used in conjunction to predict defaults [4].

2. Research Question:

1. What predictors could simulate analytics the SBA uses in a loan approval process?

2. Which prediction model is most accurate?

3. Data:

3.1 Data Description

The dataset is subsetted from the U.S.SBA loan database and records data from 1987 through 2014 (899,164 observations) with 27 variables. Some of these variables are: whether the economy was in a recession at the time, the SBA approved loan amount, the time period of the loan, whether a loan was paid off in full, or if the SME had to charge off any amount and how much that amount was.

3.2 Data Cleaning

Before exploring the data, we needed to clean the data: address the missing values, and test for possible multicollinearity between features. The dataset was removed of NA values, either by deleting observations with too many NAs or by imputing values based on the median value of the feature. Then, we used the vifstep() function to identify individual variables that might be highly correlated, identifying 9 variables from the results. To avoid high variance and unstable results from the analyses, we removed these variables from the dataset. We also checked the distribution of variables in the dataset and noticed that the variable ‘ChgOffPrinGr’ was highly skewed towards 0 with an outlier extending to 200000. Even after removing the outliers, the variable remained extremely skewed and unevenly distributed between the two categories of the response variable, and hence, we decided to remove it.

For our data analyses, our response variable is “Default”- whether the loan was paid off in full or if the SME had to charge off any amount and how much that amount was.

4. Methodology:

The first models we decided to use are the logistic regression models. Logistic Regressions explain the relationship between one dependent binary variable - in our case “Default” - and the independent variables [5]. Though tree classification methods are sometimes more accurate, we decided to also use logistic regressions for easier interpretability of the model. We believe that this model - hand in hand with a random forest - can help us understand the higher weighing factors in the prediction process.

4.1 Logistic Regression Models

The mathematical formulation for logistic regression is as follows:

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i,1} + ... + \beta_k X_{i,k}.$$

where k is the number of predictors. From our cleaned dataset, we determine that k =18. We will then use the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) to score the performance of the regression model. The final candidate models that were fit were 1) logit link using AIC stepwise procedure, 2) logit link using AIC stepwise procedure with interaction terms, 3) logit link using BIC stepwise procedure, and 4) logit link using BIC stepwise procedure with interaction terms. The optimal model obtained with the AIC criterion (Model 1) had 3 predictors and the optimal model obtained with AIC criterion including interaction terms (Model 2) had 16 predictors. The optimal model obtained with the BIC criterion (Model 3) had 3 predictors and the optimal model obtained with BIC criterion including interaction terms (Model 4) had 4 predictors.

4.2 Random Forest

The second model we built is the Random Forest model (tree classification). Random forests are developed by averaging the decision trees in the forests for prediction or regression [6]. While a random forest’s bootstrapping method decorrelates the trees, it also reduces variance by averaging the trees in the forest [7]. Forests are also known to model large datasets while avoiding over-fitting. Furthermore, we wanted to compare the model’s “best predictors” to the ones we had chosen for the logistic regression.

5. Results:

5.1 Logistic Regression

To determine the performance of our models and compare the logistic regression models with the Random Forest model, we calculated AUC scores and the performance metrics of accuracy, sensitivity, specificity, and precision for the 4 logistic regression models and misclassification rate and variable importance for the Random Forest model. The AIC model (Model 1) shows *NAICS, ApprovalFY, RealEstate1, Portion,* and *Recession1* as the most significant predictors (alpha = 0.001) [in Appendix A.1]. The AIC with interaction model (model 2) shows *NAICS, ApprovalFY, Recession1, RealEstate1:Portion, ApprovalFY:Recession1, ApprovalFY:RealEstate1* as the most significant terms [in Appendix A2]. The BIC model (Model 3) shows *NAICS, ApprovalFY, RealEstate1, Portion,* and *Recession1* as the most significant predictors [in Appendix A.3]. The BIC with interaction terms model (Model 4) shows high significance of *NAICS,ApprovalFY,RealEstate1, Recession1, RealEstate1:Portion, ApprovalFY:Recession1, ApprovalFY:RealEstate1* [in Appendix A.4].

On calculating the performance metrics, we observed that 4 logistic regression models yielded accuracy scores ranging from 0.69 to 0.716 which signifies moderate correct classification. The AUC scores for the 4 models are between 0.83 and 0.87, with the AIC with interaction model (Model 2) having the highest AUC score of 0.8716 and the highest accuracy score of 0.716. This model highlights ApprovalFY, RealEstate1, and Recession1 as being the highest significant predictors featuring in interaction terms as well.

5.2 Random Forest

For our next model, We decided to use the random forest method following logistic regression. Here, DisbursementGross had the highest measure of variable importance based on the Gini coefficient, followed by ApprovalFY, and Zip. It is also observed that the OOB error rate, which is the average error for each calculated using predictions from the trees that do not contain their respective bootstrap sample, is 21.96%. A confusion matrix after predicting the test set yields the overall misclassification rate to be 18.8%.

Conclusions and Other Considerations:

6.1 Conclusions

We concluded that out of our logistic regression models, AIC with Interaction model is best used to highlight the useful predictors for loan eligibility. These are ApprovalFY, RealEstate1, and Recession1 and their interactions. This answers the first part of our research question. To answer the second part of our research question, we choose the Random Forest model as the most accurate model in our project with a misclassification rate of 18%. The Random Forest model also highlights ApprovalFY as a variable of importance and DisbursementGross. Hence, we conclude that Random Forest is the most accurate model to use in this project and the most important predictors are ApprovalFY, DisbursementGross, RealEstate at Level 1, Recession at Level 1, and the interactions between ApprovalFY, RealEstate at Level 1, and Recession at Level 1.

6.2 Considerations

Fitting prediction models on this dataset, even if accurate, may cause bias as the response variable is imbalanced. To correct the balance of this dataset, data ethics need to be evaluated to effectively trim the dataset but also have it as representative which must be ensured in order for the model to retain validity and generalizability. Additionally, using a bigger dataset - thought would be available to the SBA - may reduce the skewness of the data, given more observations.

As next steps, we would like to use more variables in our analysis and enhance prediction accuracy. We would also like to investigate more models to build one suited for different cities in the United States with the scope to be generalized and help small businesses and banks across the world.

References:

[1] United States, Office of Advocacy. Frequently Asked Question, U.S. Small Business Administration, 2012, pp. 1–4 https://www.sba.gov/sites/default/files/FAQ_Sept_2012.pdf

[2] United States, Office of Advocacy. Frequently Asked Question, U.S. Small Business Administration, 2020, pp. 1–4 https://cdn.advocacy.sba.gov/wp-content/uploads/2020/11/05122043/Small-Business-FAQ-2020.pdf

[3] Zhu, Lin, et al. “A Study on Predicting Loan Default Based on the Random Forest Algorithm.” *Procedia Computer Science*, Elsevier, 31 Dec. 2019, https://www.sciencedirect.com/science/article/pii/S1877050919320277.

[4] Zhao, Selena, and Jiying Zou. “Predicting Loan Defaults Using Logistic Regression.” *Journal of Student Research*, Jstor, 10 Mar. 2022, https://www.jsr.org/hs/index.php/path/article/view/1326.

[5] “What Is Logistic Regression?” *Statistics Solutions*, 11 Aug. 2021, https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-logistic-regression/.

[6] Finnstats. “Random Forest in R: R-Bloggers.” *Https://Www.r-Bloggers.com/*, Www.r-Bloggers.com, 13 Apr. 2021, https://www.r-bloggers.com/2021/04/random-forest-in-r/#:~:text=Random%20Forest%20in%20R%2C%20Random,to%20identify%20the%20important%20attributes.

[7] Walia, Anish Singh. “Random Forests in R.” *DataScience+*, Https://Datascienceplus.com/, 24 July 2017, https://datascienceplus.com/random-forests-in-r/.

Appendix

A.1.AIC MODEL (Model 1)

```
Call:
glm(formula = Default ~ NAICS + ApprovalFY + NoEmp + UrbanRural +
    RevLineCr + RealEstate + Portion + Recession, family = binomial(link = "logit"),
    data = vifsubset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9999  -0.7889  -0.2561   0.8979   2.9762

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.833e+02  9.395e+01  -1.951  0.05103 .
NAICS       -7.563e-04  1.224e-04  -6.181 6.37e-10 ***
ApprovalFY   2.929e-01  2.929e-02   9.999  < 2e-16 ***
NoEmp       -1.611e-02  6.320e-03  -2.549  0.01080 *
UrbanRural1 -7.377e-01  3.482e-01  -2.118  0.03414 *
UrbanRural2 -7.776e-01  4.023e-01  -1.933  0.05326 .
RevLineCr1   3.546e-01  1.677e-01   2.114  0.03449 *
RevLineCr2  -4.643e-01  1.680e-01  -2.764  0.00571 **
RealEstate1 -2.150e+00  2.577e-01  -8.341  < 2e-16 ***
Portion     -2.274e+00  5.170e-01  -4.399 1.09e-05 ***
Recession1   9.059e-01  2.008e-01   4.511 6.46e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2651.3  on 2098  degrees of freedom
Residual deviance: 1924.4  on 2088  degrees of freedom
AIC: 1946.4
```

A.2.AIC MODEL with INTERACTION (Model 2)
(a subset of output)

```
    data = vifsubset)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.63141  -0.72604  -0.04557   0.81782   2.82824

Coefficients: (6 not defined because of singularities)
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                 4.219e+02  2.535e+04   0.017  0.98672
Zip                        -1.112e-02  2.694e-01  -0.041  0.96707
NAICS                      -6.853e-04  1.507e-04  -4.548 5.42e-06 ***
ApprovalFY                  4.926e-01  1.209e-01   4.075 4.60e-05 ***
NoEmp                      -3.618e-02  2.525e-02  -1.433  0.15186
NewExist1                   2.703e+00  1.885e+03   0.001  0.99886
NewExist2                   3.015e+00  1.455e+03   0.002  0.99835
CreateJob                   1.182e+02  5.897e+01   2.005  0.04495 *
RetainedJob                 3.503e-02  2.345e-02   1.494  0.13522
FranchiseCode              -4.931e-02  1.947e-02  -2.532  0.01134 *
UrbanRural1                 3.170e+02  2.155e+02   1.471  0.14128
UrbanRural2                 9.342e+02  2.858e+02   3.269  0.00108 **
RevLineCr1                  4.243e+02  1.540e+02   2.756  0.00586 **
RevLineCr2                 -1.576e+02  1.626e+02  -0.969  0.33232
LowDoc1                    -1.045e+03  2.530e+04  -0.041  0.96707
LowDoc2                    -1.136e+03  2.530e+04  -0.045  0.96418
DisbursementGross          -3.521e-05  1.458e-05  -2.414  0.01578 *
RealEstate1                 6.285e+02  1.323e+03   0.475  0.63464
Portion                    -7.129e-01  6.291e-01  -1.133  0.25711
Recession1                 -1.306e+03  3.287e+02  -3.972 7.12e-05 ***
RealEstate1:Portion        -2.218e+01  4.697e+00  -4.722 2.33e-06 ***
ApprovalFY:Recession1       6.510e-01  1.639e-01   3.973 7.10e-05 ***
ApprovalFY:RealEstate1      8.284e-01  2.067e-01   4.008 6.11e-05 ***
```

A.3.BIC MODEL (Model 3)

```
Call:
glm(formula = Default ~ NAICS + ApprovalFY + NoEmp + RevLineCr +
    RealEstate + Portion + Recession, family = binomial(link = "logit"),
    data = vifsubset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9477  -0.7981  -0.2501   0.9111   3.0543

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.216e+02  8.943e+01  -1.359  0.17407
NAICS       -7.675e-04  1.220e-04  -6.293 3.10e-10 ***
ApprovalFY   2.647e-01  2.631e-02  10.059  < 2e-16 ***
NoEmp       -1.620e-02  6.304e-03  -2.569  0.01020 *
RevLineCr1   3.606e-01  1.683e-01   2.143  0.03215 *
RevLineCr2  -4.530e-01  1.676e-01  -2.702  0.00688 **
RealEstate1 -2.221e+00  2.550e-01  -8.708  < 2e-16 ***
Portion     -2.043e+00  5.026e-01  -4.066 4.79e-05 ***
Recession1   9.128e-01  2.020e-01   4.520 6.19e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2651.3  on 2098  degrees of freedom
Residual deviance: 1928.8  on 2090  degrees of freedom
AIC: 1946.8

Number of Fisher Scoring iterations: 6
```

A.4.BIC MODEL with INTERACTION (Model 4)
(a subset of output)

```
glm(formula = Default ~ NAICS + ApprovalFY + RevLineCr + DisbursementGross +
    RealEstate + Portion + Recession + RealEstate:Portion + ApprovalFY:Recession +
    ApprovalFY:RealEstate + RevLineCr:RealEstate, family = binomial(link = "logit"),
    data = vifsubset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6725  -0.8066  -0.1170   0.8780   2.7895

Coefficients: (1 not defined because of singularities)
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -2.538e+01  9.185e+01  -0.276  0.78231
NAICS                   -7.839e-04  1.260e-04  -6.222 4.90e-10 ***
ApprovalFY               2.206e-01  2.580e-02   8.552  < 2e-16 ***
RevLineCr1               4.442e-01  1.779e-01   2.496  0.01255 *
RevLineCr2              -3.146e-01  1.718e-01  -1.831  0.06717 .
DisbursementGross       -8.125e-07  2.868e-07  -2.833  0.00461 **
RealEstate1             -1.535e+03  3.536e+02  -4.341 1.42e-05 ***
Portion                 -8.414e-01  5.360e-01  -1.570  0.11648
Recession1              -1.098e+03  2.700e+02  -4.064 4.82e-05 ***
RealEstate1:Portion     -1.797e+01  3.461e+00  -5.193 2.07e-07 ***
ApprovalFY:Recession1    5.481e-01  1.348e-01   4.067 4.75e-05 ***
ApprovalFY:RealEstate1   7.717e-01  1.770e-01   4.361 1.30e-05 ***
RevLineCr1:RealEstate1 -2.939e+00  9.161e-01  -3.208  0.00134 **
```

B.1. Random Forest Output

```
Call:
 randomForest(formula = Default ~ ., data = trainset, ntree = 100,      mtry = 4, na.action = na.roughfix)

               Type of random forest: classification
                     Number of trees: 100
No. of variables tried at each split: 4

        OOB estimate of  error rate: 21.96%
Confusion matrix:
    0   1 class.error
0 962 167   0.1479185
1 202 349   0.3666062
```

B.2. Random Forest Confusion Matrix

```
   pred.forest
      0   1
  0 246  40
  1  39  95
```

B.3.Random Forest Variable Importance Plot