# 

# 

# **SBA Loan Data-Driven Solution**

By: Esteban Diaz and Emmanuel Zarate

# 

# 

# 

# 

# 

# 

# 

# 

# 

# **The Problem**

Banks in the U.S. face challenges in determining whether a small business loan application is high-risk or low-risk. Some factors influence loan defaults more than others, and our goal is to provide a data-driven approach to help banks minimize loan defaults while still supporting small business growth. By analyzing key risk factors such as Gross Approved Loan (**GrAppv**), Disbursed Loan Amount (**DisbursementGross**), Loan Term (**Term**), Number of Employees (**NoEmp**), Business Existence (**NewExist**), and Low Documentation Loans (**LowDoc**), we can determine which variables contribute most to loan default risk. Our analysis provides banks with actionable insights to improve loan approval processes and reduce financial risks.

### **Data Understanding and Visualization**

To develop an effective model, we first explored the dataset to identify the most important factors contributing to loan risk. The original dataset contained extra information that was not necessary for predictive modeling, so we selected only the most relevant features, including borrower name, bank name, loan term, number of employees, whether the business already existed, participation in the Low Documentation Program (**LowDoc**), loan disbursement amount (**DisbursementGross**), loan status (**MIS\_Status**), gross approved loan amount (**GrAppv**), and SBA’s approved loan guarantee (**SBA\_Appv**). These attributes are directly related to whether a business is successful and can repay the loan.

For instance, a longer loan term or a LowDoc loan may be an indicator of higher risk, while a business with more employees may have better financial stability. To better understand the relationships between key variables, we performed data visualization using scatterplots, boxplots, and bar graphs. The scatterplot of **DisbursementGross** vs. **GrAppv** shows a positive linear relationship with the **Status** of the loan, indicating that larger loan amounts often correlate with higher default rates. Additionally, the bar graph illustrates the relationship between **LowDoc** loans and default rates. The results reveal that borrowers using LowDoc loans defaulted at a much higher rate compared to those who provided full documentation. While some businesses using LowDoc loans successfully repaid their loans, the default rate was significantly higher, making **LowDoc** a crucial predictor of loan risk.

### **Data Preparation**

To prepare the dataset for modeling, we performed data wrangling to improve accuracy and efficiency. First, we removed unnecessary variables and handled missing values by omitting incomplete cases. Next, we converted the target variable (**MIS\_Status**) into a binary format: 1 for Paid in Full and 0 for ChargeOff. We also filtered the dataset to remove rows where **LowDoc** was not Y or N, and where **NewExist** was unknown (0 value). Since the dataset was too large to process in our models efficiently, we created a smaller sample of 5,000 loans, ensuring that the data was large enough to be accurate, while giving our computers the ability to perform the models.

### 

### **Modeling**

The Logistic Regression Model used Status as the target variable and **DisbursementGross, GrAppv, Term, NoEmp, NewExist, LowDoc**, and **SBA\_Appv** as predictor variables. Our findings showed that longer loan terms (**Term**), businesses with more employees (**NoEmp**), and Low Documentation Loans (**LowDoc**) were the most significant indicators of loan default, as they had positive coefficients. Conversely, loan amount variables (**DisbursementGross, GrAppv, SBA\_Appv**) were not statistically significant predictors of default risk. The model achieved an accuracy rate of 82%, meaning it correctly predicts loan risk in around 8 out of 10 cases. The K-Nearest Neighbors (KNN) Model used Status as the target variable and DisbursementGross, GrAppv, Term, NoEmp, and SBA\_Appv as predictors (excluding NewExist and LowDoc since KNN requires only numerical variables). We determined that the optimal number of neighbors (k) was 3, which provided the highest accuracy of 86.2%. The confusion matrix for the validation set confirmed 85.6% accuracy, with a Kappa score of 0.435, indicating an agreement between predicted and actual values.

### **Evaluation and Insights**

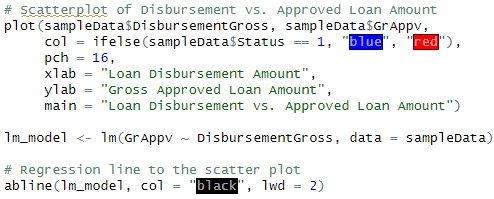
Comparing both models, KNN achieved slightly higher accuracy (86%) than Logistic Regression (82%), but Logistic Regression provides better interpretability, allowing banks to understand the weight of different risk factors. We know that banks want to find the best solution for mitigating the risk that they are taking so using the more accurate model would be our suggestion. Our findings also suggest that banks should be particularly cautious of longer loan terms, businesses with more employees, and loans under the LowDoc program, as these are the strongest predictors of loan default.

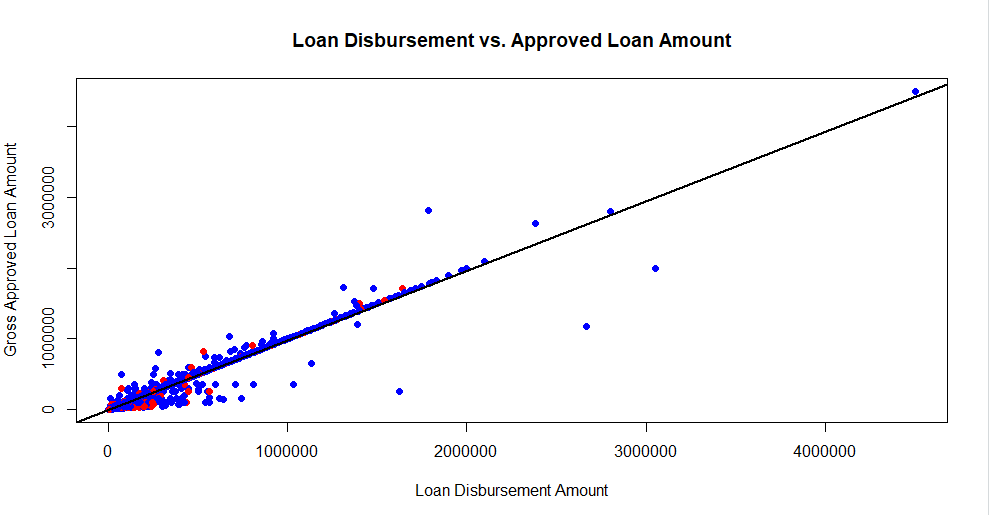
### **Conclusion**

To help banks improve loan approval decisions, we recommend implementing a data-driven risk assessment system based on our models. Banks should flag high-risk loans based on loan terms, employee count, and LowDoc status and consider implementing stricter approval criteria for loans classified as high-risk. Additional documentation should be required for LowDoc loans, as they show a much higher probability of default. Furthermore, banks should reconsider offering extended loan terms to high-risk businesses, as our models indicate that longer loan terms are strongly associated with increased default risk. By integrating our models into their decision-making processes, banks can reduce default rates while still providing financial support to businesses that demonstrate lower risk. This approach balances risk management with continued small business growth, ensuring that banks can make smarter, data-driven loan approvals in the future.

Appendix A

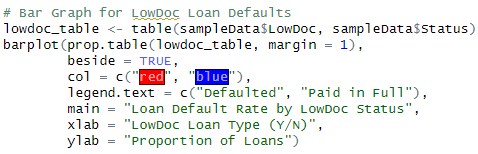
Data Understanding

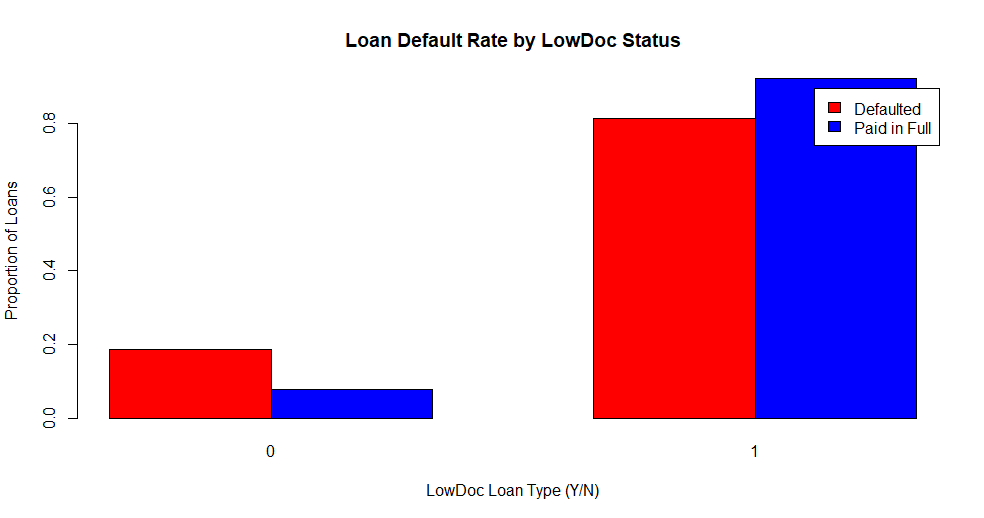




Appendix B

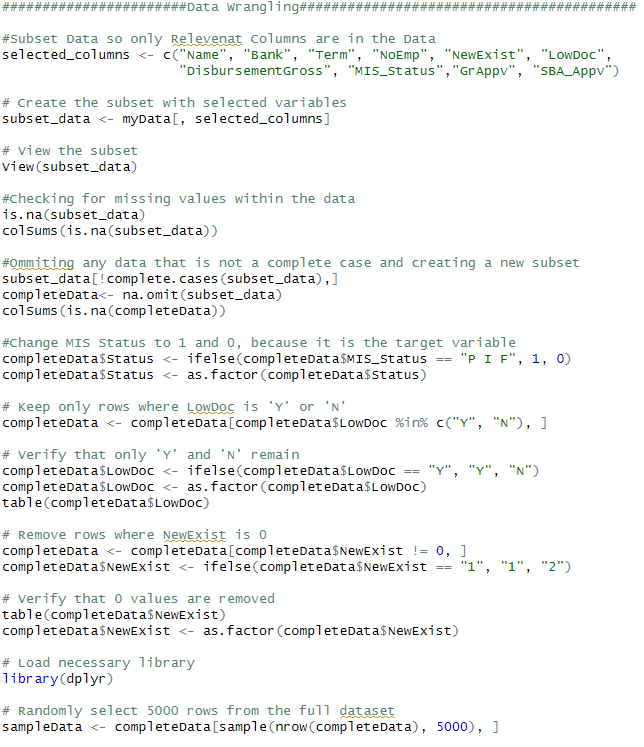
Data Understanding

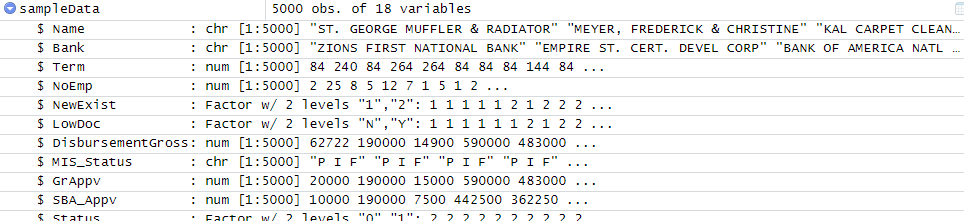




Appendix C

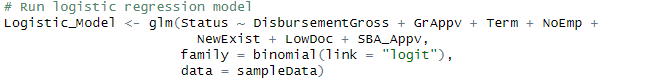
Data Wrangling

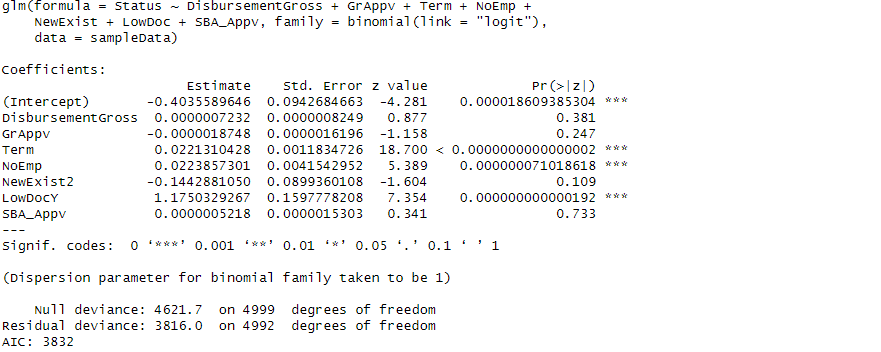


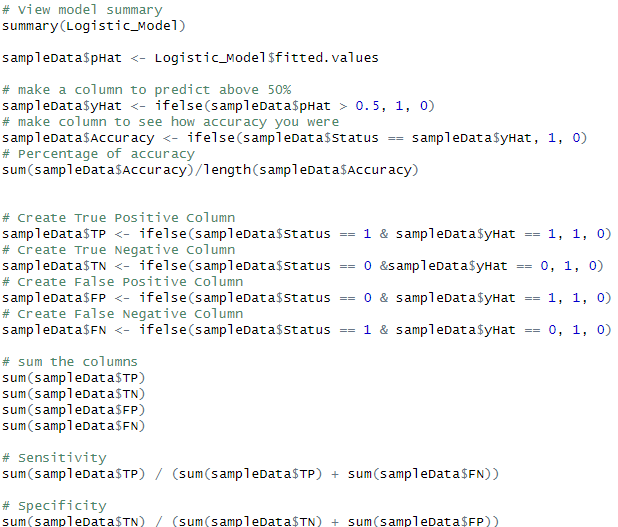


Appendix D

Logistic Regression Model

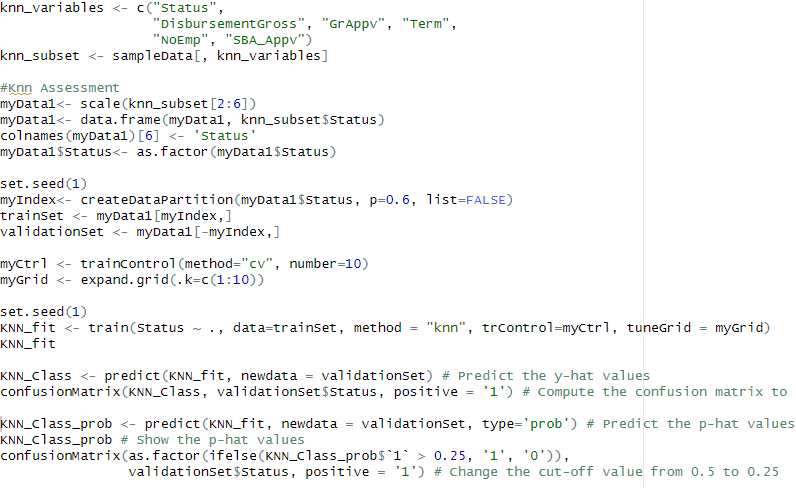


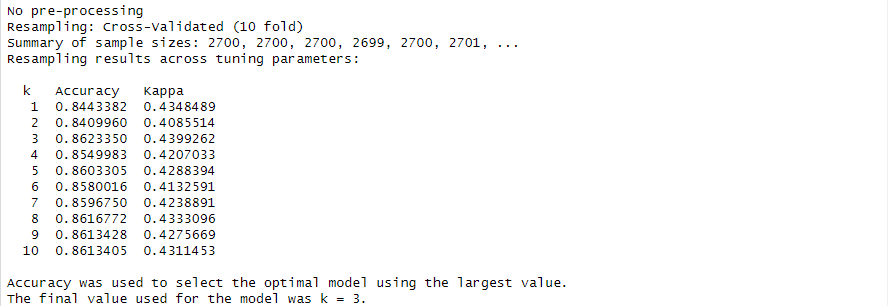




Appendix E

KNN Analysis on Train Set



****

Appendix F

Confusion Matrix on Validation Set



