Loan Prediction

Team: Jemin Patel, Dhaval Patel, Rahul Surve, Sarthak Dhaneshwar
Instructor : Chris Asakiewicz

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Introduction & Background

- Loan approval & recovery of loans is a crucial contributing aspect in the financial statements of a bank
- US banks set aside $55 Billion to protect probable loan losses and the number increased to approximately $210 billion throughout pandemic
- In Q3 2022 results, 6 major US banks reported a decline in profits. Bad loans are the main cause of this situation.
- We're attempting to solve this issue for the bank. Finding out if the applicant has the financial ability to repay the loan that has been approved for them is the main goal of our problem statement.

Data Scope

- We are leveraging historical information of loan applicants over a span of 6 months of US banks collected from GitHub
- The dataset covers numerous applicant attributes such as demographics, salary, loan terms, credit rating, etc.
- After data processing, we are using 11 independent variables and 1 dependent variable (loan status being granted or not) for our model.
- The data we are utilizing is great to anticipate if the applicant can repay the loan or not because all of the applicant's criteria have a substantial impact on whether their application will be granted or not.

Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area
Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0	Rural
Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0	Urban
Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0	Urban
Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0	Urban
Male	Yes	2	Graduate	Yes	5417	4196.0	267.0	360.0	1.0	Urban

Independent variables

Loan_Status
N
Y
Y
Y
Y

Dependent variable

Methodology

- The problem that we are seeking to solve falls under the field of predictive analytics.
- As our predicted outcomes were binary (Yes - Applicant should be allowed loan; No - Applicant should not be granted loan), we initially employed a Logistic Regression model.
- Later, we implemented ensemble modeling and used a variety of machine learning techniques, including K-Nearest Neighbor, Support Vector Machine, and Gradient Boost.

Model Accuracy & Result

- The model was tested using 25% of the data, with the remaining 75% being utilized for training. Among all the algorithms, the Logistic Regression Model had the best accuracy.

```
shape of training data :  (360, 24)
shape of testing data :  (120, 24)
              precision    recall  f1-score   support

           0       0.93      0.34      0.50        38
           1       0.76      0.99      0.86        82

    accuracy                           0.78       120
   macro avg       0.85      0.66      0.68       120
weighted avg       0.82      0.78      0.75       120

[[13 25]
 [ 1 81]]
LR accuracy: 78.33%
```

- In this case, there is just one false negative, meaning that the model correctly identified all other applicants as being qualified. One candidate for the bank missed out on the opportunity cost as a result of this.
- Our algorithm correctly predicted that 81 applicants should not receive loans, while it incorrectly projected that 25 applications should. This indicates that three out of every four occasions, our model will correctly identify a loan application who should not be approved.

```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        38
           1       0.68      1.00      0.81        82

    accuracy                           0.68       120
   macro avg       0.34      0.50      0.41       120
weighted avg       0.47      0.68      0.55       120

[[ 0 38]
 [ 0 82]]
SVC accuracy: 68.33%
```

- There is not a single false negative value in this SVM model, indicating that no excellent candidates were predicted to be terrible applicants. In essence, this implies that no business opportunity was missed.
- However, when compared to a logistic regression model, the model's overall accuracy is far too low. Loss brought on by bad loans is far greater than missed business opportunities. So, for our model, the logistic regression approach is far more successful.

Business Impact

- The average loan amount is $140,000, the average loan period is 30 years, and the average interest rate is 5% each year.
- When we compare these parameters to the results of our model's confusion matrix, we get:

1. Only one candidate was wrongfully denied, therefore the opportunity cost was minimal. As a result, the bank only lost $1,33,216.
2. Bad Loans - After using our model, the Non-Performing Loan Ratio of our dataset decreased from 0.4 to 0.25. Now, for the purpose of consistency, we are considering merely an average loan amount. As a result, the bank was able to save $7,717,500 using our approach.