Overview of Analysis

Lending institutions extend funds or assets to borrowers with the anticipation that borrowers will either return the asset or reimburse the lender. Credit Risk arises when a borrower fails to return an asset or repay a loan, resulting in financial loss for the lender. Lenders assess this risk through various methods, yet in this analysis, we'll utilize Machine Learning to examine a dataset of past lending transactions from a peer-to-peer lending service. The aim is to construct a model capable of gauging the creditworthiness of borrowers.

Employing a machine learning model, I aim to discern between low-risk (healthy) and high-risk (non-healthy) loans based on the loan status provided by the lending institution.

Opting for the Logistic Regression Algorithm seems most fitting for our machine learning approach, given its widespread application in predicting the probability of a target variable in classification scenarios.

Utilizing the lending company's dataset, I constructed a Logistic Regression Model that achieved an accuracy score of 95%. Despite the model's high accuracy, its recall value for non-healthy loans stands at 0.91, slightly lower than the 0.99 recall value for healthy loans. This discrepancy suggests the model's inclination to predict loan statuses as healthy rather than as non-healthy. This bias is attributed to the dataset's imbalance, where one class label (in this case, healthy loans) substantially outweighs the other (non-healthy loans).

Below is a screenshot of Step 3 where we can see that the data is highly imbalanced by using the value\_counts function. The majority is healthy loans (0) while the minority is non-healthy loans (1).

A screenshot of a computer

Description automatically generated

Under Step 3: Evaluating model’s performance – Confusion Matrix: We found that among the 18,765 loans categorized as healthy (low-risk), the model accurately identified 18,663 as healthy and misclassified 102 as healthy. Similarly, out of the 619 loans labeled as non-healthy (high-risk), the model correctly classified 563 as non-healthy and inaccurately identified 56 as non-healthy.

A screenshot of a computer code

Description automatically generated

To improve the accuracy score and enhance the model's ability to detect misclassifications in non-healthy loans, we can employ the RandomOverSampler module from the imbalanced-learn library. This technique involves oversampling the data by adding additional instances of the minority class (non-healthy loans) to achieve a more balanced dataset.

A screenshot of a computer code

Description automatically generated

Leveraging the lending company's dataset, I developed a Logistic Regression Model trained on oversampled data, yielding an accuracy score of 99%, surpassing the performance of the model trained on imbalanced data. The improved performance of the oversampled model can be attributed to the dataset's balance. Notably, the recall value for non-healthy loans escalated from 0.91 to 0.99, signifying the model's remarkable proficiency in identifying errors, specifically in labeling non-healthy (high-risk) loans as healthy (low-risk).

Results

The Logistic Regression model trained on the Imbalanced DataSet achieved perfect accuracy in predicting healthy loans (low-risk) and an 85% accuracy in predicting non-healthy loans (high-risk). Inherent in the model trained on imbalanced data are higher probabilities of these errors occurring:

1. Misclassifying a healthy loan (low-risk) as a non-healthy loan (high-risk).
2. Misclassifying a non-healthy loan (high-risk) as a healthy loan (low-risk).

Evaluating the model's recall scores reveals that it made 1% errors in predicting healthy loans and 9% errors in predicting non-healthy loans.

**Machine Learning Model 1: Evaluation of Model 1's Accuracy, Precision, and Recall scores.**

Relative to the original dataset, the count of healthy loans exceeds that of unhealthy loans. The model exhibits an impressive accuracy of 99%, boasting a perfect precision score of 100% for class 0 (representing healthy loans) and a respectable 85% precision for class 1 (unhealthy loans). Additionally, it demonstrates noteworthy recall rates, achieving 99% for predictions of class 0 and 91% for identifying high-risk loans categorized under class 1.

**Machine Learning Model 2: Analysis of Model 2's Accuracy, Precision, and Recall scores.**

The accuracy score for this model also stands at an impressive 99%. Upon reviewing the confusion matrix, the oversampled data model notably excels in predicting false negatives, misidentifying only 4 loans of type 0. Similar to the prior model, precision remains perfect at 100% for type 0 loans and reaches 84% for type 1 loans. Notably, there's an improvement in the recall score for high-risk loans compared to the previous model.

Summary

A lending institution seeks a model capable of accurately discerning between healthy and non-healthy loans, aiming to mitigate potential financial repercussions. Misclassifying healthy loans as non-healthy could lead to customer loss, while mislabeling non-healthy loans as healthy may result in financial losses for the company.

The Logistic Regression model, trained using oversampled data, outperformed its imbalanced data counterpart. The adoption of a balanced dataset approach notably enhanced both accuracy and recall, reducing errors in categorizing non-healthy loans.

In an effort to minimize risks, the lending company prioritizes minimizing false positives, where non-healthy loans are erroneously classified as healthy. A comparative examination of confusion matrices reveals the following:

Machine Learning Model 1 with imbalanced data:

* 56 false positives (actual value: healthy, predicted value: non-healthy)
* 102 false negatives (actual value: non-healthy, predicted value: healthy)

Machine Learning Model 2 with balanced data:

* 4 false positives (actual value: healthy, predicted value: non-healthy)
* 116 false negatives (actual value: non-healthy, predicted value: healthy)

Given these findings, the model utilizing balanced data is strongly recommended due to its significant reduction in false positives and improved accuracy in distinguishing between healthy and non-healthy loans.